Title: Evolution-based Functional Genomics
Inventor: Steven A. Benner
Citizenship: USA
Permanent Address: 1501 N.W. 68th Terrace, Gainesville FL 32605
Tel: (352) 332 2262
Fax: (352) 846 2580

Cross Reference To Related Applications

This application is a continuation-in-part of application Serial No. 07/857,224, filed March 25, 1992, and issued as US Patent 05958784 on September, 28, 1999, and application Serial No. 08/914,375 filed August 08, 1997, currently pending.

102/20

Statement of rights to inventions made under Federally-sponsored research: None

INTRODUCTION

Field of the Invention:

This invention relates to the area of bioinformatics, more specifically to methods for analyzing the sequences of evolutionarily related proteins, and most specifically for identifying evolutionary and functional relationships between proteins and the genes that encode them.

Background

Proteins are linear polypeptide chains composed of 20 different amino acid building blocks. Determining the sequence of amino acids in a protein is now experimentally routine, both by direct chemical analysis of the proteins themselves, and by translation of genes that encode proteins. Genome sequencing projects provide those genes in abundance, and the full complement of genes is known for many microorganisms, yeast, the worm *C. elegans*, and the fly *Drosophila melanogaster*. The size of protein sequence databases will grow explosively over the next decade as more genome sequencing projects are completed.

While genomic sequence data are widely believed to hold the key to a revolution in biology, and while their impact on experimental work is already being felt, much of the revolution has still not materialized. Missing in particular are bioinformatics tools that extract information about biological function starting from genomic data and existing biological information, in a form that can be used by biomedical researchers. This makes it difficult to asign a "utility" to a gene sequence, in turn making it difficult to capture commercial value from gene sequence data. This issue is found throughout the commerce of modern genomics, including in the patenting of genes. It is still not clear what standard must be applied to establish that a gene patent application meets the "utility" requirement of the code, but information concerning the biological function of the gene for which a patent is sought is almost certainly helpful. Tools that generate testable hypotheses about the biological function of genes that encode proteins have utility precisely because they enable genome sequences to be exploited commercially.

Once those tools are in place, it should be possible to mine genomic sequence data for information about pathways [Mar99], generate insights into complex biological function (e.g. development) [Rub00], and perhaps even identify pharmaceutical targets [Pol00]. Clearly, the payoff from a tool that could provide biological information from sequence data could be enormous. Experiments are expensive; genomic data are not (or, more precisely, their cost has already been amortized).

For this reason, functional bioinformatic tools need not be generate "indisputable conclusions" for them to have a broad utility. The payoff would be substantial if the tool did nothing more than suggest a hypothesis about the function of a gene, or rule out some fraction of a protein family as relevant for a function, both ways of targeting subsequent experiments. Modern biomedical research, especially that directed towards the development of therapeutic agents, is largely a random walk in any case. Functional bioinformatic tools would be valuable if they do nothing more than favorably bias that walk.

Generating hypotheses concerning possible biological functions that a gene might have is part of the "annotation problem" in modern genomics. "Annotation", and (indeed) "function", mean very different things to different people. Much of the literature regards "function" as equivalent to "behavior", what is measured in the laboratory. Crystallographers occasionally view "function" as equivalent to "structure". Throughout this proposal, we distinguish between "homology" (relationship by common ancestry), "structure" (at two levels, as known to those skilled in the chemical arts, "constitution", meaning "sequence", and "conformation", which is commonly referred to as "the fold" or, incorrectly, "the structure" by structural biologists), "behavior" (what is measured in the laboratory), and "function". Under Darwinian theory, "function" refers to adaptive behavior, properties that confer fitness, the ability of an organism to survive and reproduce. The Darwinian paradigm holds that the only way to achieve function is by random variation of genetic structure (mutation) followed by natural selection.

It is important to keep these distinctions clear. A statement concerning the function of a gene is ultimately a statement about how that gene contributes to the fitness of the host organism. For this reason, homologous proteins in different species generally do not have "the same function", as different species have different requirements for fitness. They may, however, have "analogous" functions. As we disclose below, even subtle differences in function between orthologous proteins in different organisms can be very interesting, and can be the key to delivering a true understanding of "function" to biological and biomedical research scientists. Therefore, tools that suggest (again, as a hypothesis) that the function of two homologous genes might be different are frequently as useful as those that suggest that the two genes have analogous functions.

In any case, the fact that Darwinian processes have generated the genes through natural selection acting on the encoded proteins means that an evolutionary analysis must at some point be involved in any annotation procedure; an evolutionary analysis is necessary to analyze for function. Further, as we shall disclose, an evolutionary analysis is sufficient to generate hypotheses concerning function. As with any hypothesis in science, functional hypotheses cannot be "proven". Rather, as with any hypothesis in science, they form part of a web of conjecture, hypothesis, data, and interpretation that builds a useful picture of function; utility comes long before "proof", in part because"proof" is not possible for any substantive statement about the natural world.

The most common way in which evolutionary analysis is used today to annotate sequences involves pairwise sequence comparisons to detect homology between two proteins. The conventional recipe for inferring the "function" of a target open reading frame (the target sequence) using evolutionary analysis follows five steps:

1. Use the target sequence as a probe in a BLAST [Alt90] or FASTA [Lip85] search of the Genbank database (or an equivalent).
2. Identify "hits", proteins in the database whose sequence resembles that of the target.
3. Evaluate the hits based on a statistical model."
4. Download the annotation of the statistically best "hits" that have functional annotation of their own.
5. Infer that the function of the target protein is the same as the function of the best protein hit

Those annotating genomic sequences using this recipe recognize several obvious limitations to the pairwise analysis approach to annotation. Commonly encountered problems include:

(a) The BLAST server returns no hits.
(b) The BLAST server returns sequences of possible homologs, but with similarity scores too low to be certain that the sequence found is indeed a homolog.
(c) The homologous sequences that the server returns have no annotation indicating function.

The consequences of these problems are mentioned in most contemporary reviews of genomics. Because of the limitations of this recipe, some 40% (depending on the details of the homology search) of the proteins in a typical genomic sequence have not been reliably assigned *any* function. It is well recognized that this arises because as powerful as it is, BLAST cannot detect homologs after their sequences have diverged far into the "twilight zone" of sequence similarity, defined (arbitrarily) by Doolittle as less than 20% sequence identity [Doo87]. For this reason, tools that detect more distant homology are being actively sought.

To solve this "no interesting hit" problem, many workers have attempted to extend the power of the pairwise alignment tool to detect increasingly distant protein sequences. Pearson for example recently developed statistical parameters from the distribution of similarity scores from thousands of unrelated sequences to gain an estimate of the statistical significance that can be used to infer sequence homology [Pea98]. BLAST has been improved by transformation into PSI-BLAST [Alt97].

Other approaches have been based on the fact that proteins diverging under functional constraint can retain their core folded structure long after sequence similarity has vanished [Ros75]. This makes possible the detection of distant homology by comparison of predicted structures. In Serial No. 07/857,224, filed

March 25, 1992, and issued as US Patent 05958784 on September, 28, 1999, a partent of this application, we disclosed the first useful tools to apply this approach. Many others have followed our lead, including a particularly interesting study by Barton and his coworkers [Rus96], and we have recently published a review article showing various successes of this approach [Ben97].

Through their ability to detect distant homologs of proteins whose functions are known, the tools disclosed in Serial No. 07/857,224 have proven to be quite useful because they generate hypotheses about proteins with unknown function. For example, these tools were applied to the heat shock protein 90 (HSP90) family, for which no member had an assigned function. A model for the conformation of the protein was built for HSP90 as part of the CASP2 prediction contest [Ger97]. The conformation model was recognizably similar to the fold for the N-terminal ATP binding domain of gyrase. This generated the hypothesis that HSP90 and gyrase were distant homologs. This generated the hypothesis that HSP90 bound ATP as it contributed to the fitness of its host organisms. This hypothesis contradicted experimental papers available at the time claimed that HSP90 did not bind ATP [Jac96]). The prediction was correct; the experimental papers were incorrect [Pro97]. And this success was recognized both by the CASP2 judges and the group that solved the crystal structure of HSP90, who wrote:

> "The tertiary fold of Hsp90 N-domain has a remarkable and totally unexpected similarity to the N-terminal ATP-binding fragment of ... DNA gyrase B protein. This similarity was not initially recognized by the authors of either the human or yeast structures but was determined [by Gerloff and Benner] within the CASP2 structure prediction competition. Our observation of specific ADP/ATP binding to Hsp90 completely contradicts the careful and widely accepted biochemical analysis of Jakob et al. (1996) who demonstrated that Hsp90 could not be photolabelled with 8-azidoATP, was not retained on C8 agarose, and did not enhance the fluorescence of MABA-ADP." [Pro97]

A year later, the tools disclosed in Serial No. 07/857,224 were used to analyze functional and structural relationships for ribonucleotide reductase [Tau97]. Other examples in our laboratory and elsewhere are summarized in [Ben97]. These examples showed how very distant homolog detection had utility in annotating gene sequences, simply by generating experimentally testable hypotheses concerning their physiological function.

The pairwise alignment approach to annotation has become dominant in contemporary genomics. Indeed, many patent applications for individual gene sequences are being submitted to the PTO where the sole argument for utility is based on the statement that the gene being patented is homologous to gene X, that homologous genes have analogous functions, and that this implies that the gene being patented and gene X have the same function. The logic here is outlined below:

(a) Sequence similarity between two proteins implies that the two proteins are homologous (share common ancestry).

(b) Homologous proteins have analogous conformations (folds).

(c) Analogous folds implies analogous behaviors (what is measured in the laboratory). Thus, homologous proteins bind analogous ligands, catalyze analogous reactions, and have analogous physical properties.

(d) Homologous proteins have analogous function, that is, they contribute to survival in analogous ways.

These assumptions constitute the "homology implies analogous function" (HIAF) logic. It is widely regard as the foundation for annotation (see, for example, Skolnik's contribution in this area [Fet98]).

Element (a) has sound statistical basis, at least within the context of a particular evolutionary theory. Element (b) is known from empirical analysis to be generally true, provided that the two proteins have diverged under functional constraints. Elements (c) and (d) are, however, are certainly not universally true, and may not be true in general. Frequently, *new* function is generated in biological systems by *recruitment* of a protein that performs a different function.

Already in 1988, well before the age of the genome, it was well known that the logic otlined above was not reliable, because of frequent recruitment in the biological world. The Applicant himself reviewed examples where conventional logic would provide deceptive annotation [Ben88]. Four examples illustrate how severely elements (c) and (d) of the logic can be violated.

The first is chosen from eubacterial enzymology, and relates to three enzymes playing three distinct roles in microbial metabolism, fumarase (in the citric acid cycle), aspartase (involved in amino acid degradation), and adenylosuccinate lyase (essential for nucleic acid biosynthesis). The three proteins are clearly recognizable as homologs. Their sequences share statistically significant similarity, as illlustrated for the following three excerpts:

LPENEPGSSIMPGKVNPTQC	fumarase
LPELQAGSSIMPAKVNPVVP	aspartase
FEKDQIGSSAMPYKRNPMRS	adenylosuccinate lyase

The overall folded form of the three proteins is the same (an 8-fold alpha-beta barrel). They catalyze reactions that, at least at a mechanistic level, have some degree of analogy. From a biologist's perspective, however, they have very different functions. The HIAF logic used by virtually every genome annotation tool would be deceived by this family.

The second example is from metazoan biology, and involves the family of proteins known as the src homology 2 (SH2) domains. SH2 domains are clearly all homologous. The proteins all have analogous folds and analogous behaviors; they all bind to a polypeptide that carries a phosphotyrosine. But they bind different peptide sequences flanking the phosphotyrosine. For this reason, they have very different functions, *as the biologist defines it* (and as it is defined throughout this proposal, the behavior that confers survival value). Some SH2 domains are in viruses, and regulate viral growth. Some participate in the immune response. Others are involved in the regulation of division of non-immune cells. For virtually any practical purpose (pharmaceutical target identification, for example), the analogies between the behaviors of different SH2 domains are less important than the differences in their function.

The third example used covarion behavior to detect changing function in elongation factors, proteins whose sequences are so highly conserved that there is little difficulty in recognizing homologs, even in the three kingdoms of life. If any protein "has the same function" in different species in orthologous form, it is elongation factors, we reasoned. Even so, covarion behavior clearly showed that different (presumably) orthologous forms of the protein had subtly different behaviors that indicate subtle differences in function. Coupling the evolutionary insights based on a sophisticated, mathematically detailed, evolutionary analysis to structural biology even identified residues that were important for these differences.

The fourth compares protein serine kinases and protein tyrosine kinases. These families are clearly homologous, the latter having been recruited from the former ca. 600 million years ago. The chemist would say that both classes of enzyme operate via analogous reaction mechanisms, differing only in the source of the oxygen nucleophile in the phosphoryl transfer reaction. The biologist would note, however, that the physiological function of the two classes of proteins are greatly different. For any biomedical application, the biologist would be correct. The physiologically relevant differences in behavior, central to the understanding of biological function (phosphorylation on tyrosine versus phosphorylation on serine) cannot be inferred for one family from the other using the conventional logic.

If proteins with recognizable (indeed, often high) sequence similarity can have different functions, the *a fortiori* argument tcan be made that surely recruitment is possible for protein homologs with marginally significant sequence similarities. This argument suggest that the focus of efforts, including those disclosed in Serial No. 07/857,224, might add only some to our ability to provide reliable annotation.

In reality, both approaches are needed. This application focuses on tools to detect and/or rule out recruitment within a protein family. We expect these tools to become increasingly more important as the genomes of metazoans are sequenced. It is now clear that the last 500 million years of molecular evolution in higher organisms has involved repeated recruitment of existing folds to perform new functions.

SUMMARY OF THE INVENTION

As discussed in Serial No. 08/914,375, the parent for the instant application, the physiological function of a biomolecule is ultimately determined by the contribution that the biomolecule makes to the efforts of the host organism to survive, select a mate (in higher organisms), and reproduce. Determining the physiological function of a protein is not trivial, however. Difficulties in establishing physiological function are discussed at length by Benner and Ellington [Benner, S. A., Ellington, A. D. Interpreting the behavior of enzymes. Purpose or pedigree? *CRC Crit. Rev. Biochem.* 23, 369-426 (1988)]. Still more difficult is identifying which behaviors of a protein as measured *in vitro* are relevant for physiological function *in vivo*. Nevertheless, the identification is important. *In vitro* behaviors that have relevance to physiological function *in vivo* are those that are interesting to study for biotechnological, biomedical, or other applications. There is at present in the art no general method for determining what *in vitro* behaviors are relevant to *in vivo* function. Processes for determining these behaviors were claimed in the parent application (Serial No. 08/914,375).

A method for making a model for the folded structure of a set of proteins from an evolutionary analysis of a set of aligned homologous protein sequences was claimed in Serial No. 07/857,224. The instant application concerns methods for using these models. The first method is used to confirm or deny a hypothesis that two proteins are homologous, and is comprised of comparing a predicted structure model for one family of proteins with a predicted structure model for a second family of proteins, or an experimental structure for the second family, and deducing the presence or absence of homology based on the presence or absence of structural similarity flanking key residues in the polypeptide sequence. The second method identifies mutations during the divergent evolution of a protein sequence that are potentially adaptive by identifying episodes during the divergent evolution of a family of proteins where there is a high absolute rate of amino acid substitution, or a high ratio of non-silent substitutions to non-silent substitutions. Amino acids that are changing during this episode are likely to be adaptive. The third is a method for identifying specific *in vitro* properties of the protein that are likely to play a physiological role *in vivo* in an organism. This methods involves synthesizing in the laboratory proteins having the reconstructed amino acid sequences of a protein before and after a period of rapid sequence evolution that characterizes adaptive substitution, measuring the *in vitro* properties of the protein before the episode of rapid sequence evolution, and then measuring the *in vivo* properties of the protein after the episode of rapid sequence evolution. The *in vitro* behaviors that remained unchanged through this episode are not likely to have adaptive significance physiologically. The *in vitro* behaviors that changed through this episode are likely to have adaptive significance physiologically. The fourth concerns method for organizing genome sized sequence databases.

BRIEF DESCRIPTION OF THE DRAWINGS

Drawing 1. The three elements modeling the evolutionary history of the leptins, proteins from the "obesity gene" identified by genetics experiments in mice. Homologs are found in other mammals (including human). (a) An evolutionary tree showing the pedigree of each leptin family member. (b) A part of the multiple alignment, showing the genetic relationship of amino acids in the protein sequence. The reconstructed ancestral sequence from the (now extinct) ancestor of humans, rodents, and ruminants (marked "X") is shown in the alignment. The sequence as shown here is deterministic; in the work to be performed here, the ancestral sequences are all probabilistic (see text)

```
.         080        090        100        110        120
     .    |     .    |     .    |     .    |     .    |
RNVIQISNDLENLRDLLHVLAFSKSCHLPWASGLETLDSLGGVLEASGYS      human
RNMIQISNDLENLRDLLHVLAFSKSCHLPWASGLETLDSLGGVLEASGYS      chimp
RNMIQISNDLENLRDLLHVLAFSKSCHLPWASGLETLDSLGGVLEASGYS      gorilla
RNVIQISNDLENLRDLLHVLAFSKSCHLPWASGLETLDRLGGVLEASGYS      orangutan
RNVIQISNDLENLRDLLHLLAFSKSCHLPLASGLETLESLGDVLEASLYS      rhesus

QNVLQIAHDLENLRDLLHLLAFSKSCSLPQTRGLQKPESLDGVLEASLYS      rat
QNVLQIAHDLENLRDLLHLLAFSKSCSLPQTRGLQKPESLDGVLEASLYS         rat
QNVLQIANDLENLRDLLHLLAFSKSCSLPQTSGLQKPESLDGVLEASLYS         mouse

RNVIQISNDLENLRDLLHLLASSKSCPLPQARGLETLESLGGVLEASLYS         ancestor X

RNVIQISNDLENLRDLLHLLASSKSCPLPQARALETLESLGGVLEASLYS         pig
RNVIQISNDLENLRDLLHLLAASKSCPLPQVRALESLESLGVVLEASLYS         sheep
RNVVQISNDLENLRDLLHLLAASKSCPLPQVRALESLESLGVVLEASLYS         ox

RNVVQISNDLENLRDLLHLLASSKSCPLPRARGLETFESLGGVLEASLYS         dog
```

Drawing 1. xxx Evolutionary tree showing the evolutionary history of the leptins. Heavy lines show branches with expressed/silent ratios higher than 2. Hatched lines show branches with expressed/silent ratios from 1 to 2. Dotted lines show branches with expressed/silent ratios less than 1, or indeterminate. Numbers on the lines indicate the ratio of expressed/silent changes for that branch. An "x" at the end of a branch signifies that a sequence for the protein is available in the database. Accoprding to the method of the instant invention, a correlation between the episode of high sequence evolution and the

Drawing 2. Evolutionary tree showing the evolutionary history of the leptin receptors. Heavy lines show branches with expressed/silent ratios higher than 2. Hatched lines show branches with expressed/silent ratios from 1 to 2. Dotted lines show branches with expressed/silent ratios less than 1, or indeterminate. Numbers on the lines indicate the ratio of expressed/silent changes for that branch. An "x" at the end of a branch signifies that a sequence for the protein is available in the database.

Drawing 3. An example of homoplasy taken from the evolution of alcohol dehydrogenase from yeast (position 30). At at least three points in the tree, a P->A substitution occurred independently.

Drawing 4. For 17 vertebrate aromatases, an unrooted evolutionary tree built by a Darwin-based (Gonnet & Benner, 1991) based on an analysis of amino acid sequences. Numbers on the branches are the K_a/K_s ratios evaluated using the methods of Fitch (1971) to reconstruct intermediate evolutionary states and Li et al. (1985). The key is given below, togetherwith the multiple sequence alignment used to calculate the tree.

1. *Tilapia nilotica* (rainbow trout), GenBank g1613859, mRNA (Chang et al., 1997)
2. *Oryzias latipes* (medaka), GenBank g1786171, ovarian follicle mRNA (Tanaka et al., 1995)
3. *Danio rerio* (zebrafish), GenBank g2306966 aromatase mRNA
4. *Carassius auratus* (goldfish) ovary, GenBank g2662330, ovarian mRNA
5. *Ictalurus punctatus* (channel catfish), GenBank g912802 (Trant, 1994)
6. *Carassius auratus* (goldfish) brain, GenBank g2662328, brain mRNA
7. *Sus scrofa* (pig) placental, isoform 2, GenBank g1762232, mRNA (Choi et al., 1997a)
8. *Sus scrofa* (pig) embryo, isoform 3, GenBank g1244543, mRNA (Choi et al., 1996)
9. *Sus scrofa* (pig) ovary, isoform 1, GenBank g1928957, mRNA (Conley et al., 1997)
10. *Bos taurus* (ox), GenBank g665546, mRNA (Hinshelwood et al., 1993)
11. *Equus caballus* (horse), GenBank g2921277, mRNA (Boerboom et al. 1997)
12. *Mus musculus* (mouse), GenBank g3046857, mRNA (Terashima et al. 1991)
13. *Rattus norvegicus* (rat), GenBank g203804, mRNA (Hickey et al., 1990)

14. *Oryctolagus cuniculus* (rabbit), GenBank g1240042, mRNA (Delarue et al, 1996)
15. *Homo sapiens* (human), GenBank g28846, mRNA (Harada, 1988)
16. *Gallus gallus* (chicken), GenBank g211703 (McPhaul et al., 1988)
17. *Poephila guttata* (zebra finch), GenBank g926845, ovary mRNA (Shen et al., 1994)

```
         010       020       030       040       050       060       070       080
           |         |         |         |         |         |         |         |
 1 MVLEMLNPMHYKVTSMVSEVVPFASIAVLLLTGFLLLVWNYKNTS-SIPGPGYFLGIGPLISYLRFLWMGIGSACNYYNK
 2 MFLEMLNPMQYNVTIMVPETVTVSAMPLLLIMGLLLLIWNCESSS-SIPGPGYCLGIGPLISHGRFLWMGIGSACNYYNK
 3 MILEMLNPMHYNLTSMVPEVMPVATLPILLLTGFLFFVWNHEETS-SIPGPGYCMGIGPLISHLRFLWMGLGSACNYYNK
 4 VLELLMQGAHNSSYGAQDNVCGAMATLLLLLLCLLLAIRHHWTEKDHVPGPCFLLGLGPLLSYCRLIWSGIGTASNYYNS
 5 -MEEVLKGTVNFAATVQVTLMALTGTLLLILLHRIFTAKNWRNQS-GVPGPGWLLGLGPIMSYSRFLWMGIGSACNYYNE
 6 VVDLLIQRAHNGTERAQDNACGATATILLLLLCLLLAIRHHRPHKSHIPGPSFFFGLGPVVSYCRFIWSGIGTASNYYNS
 7 MVLEMLNPMYYKITSMVSEVVPFASIAVLLLTGFLLLLWNYENTS-SIPSPGYFLGIGPLISHFRFLWMGIGSACNYYNE
 8 ----LVSIAPNTTVGLP-SGIPMATRSLILLVCLLLMVWSHSEKK-TIPGPSFCLGLGPLMSYLRFIWTGIGTASNYYNN
 9 MVLEMLNPMN--ISSMVSEAVLFGSIAILLLIGLLLWVWNYEDTS-SIPGPGYFLGIGPLISHFRFLWMGIGSACNYYNK
10 MVLEMLNPMHFNITTMVPAAMPAATMPILLLTCLLLLIWNYEGTS-SIPGPGYCMGIGPLISYARFLWMGIGSACNYYNK
11 VMEILLREARNGTDPRYENPRG-ITLLLLLCLVLLLTVWNRHEKKCSIPGPSFCLGLGPLMSYCRFIWMGIGTASNYYNE
12 MVLETLNPLHYNITSLVPDTMPVATVPILILMCFLFLIWNHEETS-SIPGPGYCMGIGPLISHGRFLWMGVGNACNYYNK
13 ----VVARSLCDLKCHPIDGISMATRTLILLVCLLLVAWSHTDKK-IVPGPSFCLGLGPLLSYLRFIWTGIGTASNYYNN
14 MLLEVLNPRHYNVTSMVSEVVPIASIAILLLTGFLLLVWNYEDTS-SIPGPSYFLGIGPLISHCRFLWMGIGSACNYYNK
15 MVLEMLNPIHYNITSIVPEAMPAATMPVLLLTGLFLLVWNYEGTS-SIPGPGYCMGIGPLISHGRFLWMGIGSACNYYNR
16 --------------------MPVATVPIIILICFLFLIWNHEETS-SIPGPGYCMGIGPLISHGRFLWMGVGNACNYYNK
17 MFLEMLNPMHYNVTIMVPETVPVSAMPLLLIMGLLLLIRNCESSS-SIPGPGYCLGIGPLISHGRFLWMGIGSACNYYNK

         090       100       110       120       130       140       150       160
           |         |         |         |         |         |         |         |
 1 TYGEFIRVWIGGEETLIISKSSSVFHVMKHSHYTSRFGSKPGLQFIGMHEKGIIFNNNPVLWKAVRTYFMKALSGPGLVR
 2 MYGEFMRVWISGEETLIISKSSSMFHVMKHSHYISRFGSKRGLQCIGMHENGIIFNNNPSLWRTIRPFFMKALTGPGLVR
 3 MYGEFVRVWISGEETLVISKSSSTFHIMKHDHYSSRFGSTFGLQYMGMHENGVIFNNNPAVWKALRPFFVKALSGPSLAR
 4 KYGDIVRVWINGEETLILSRSSAVYHVLRKSLYTSRFGSKLGLQCIGMHEQGIIFNSNVALWKKVRTFYAKALTGPGLQR
 5 KYGSIARVWISGEETFILSKSSAVYHVLKSNNYTGRFASKKGLQCIGMFEQGIIFNSNMALWKKVRTYFTKALTGPGLQK
 6 KYGDIVRVWINGEETLILSRSSAVYHVLRKSLYTSRFGSKLGLQCIGMHEQGIIFNSNVALWKKVRAFYAKALTGPGLQR
 7 MYGEFMRVWIGGEETLIISKSSSVFHVMKHSHYTSRFGSKPGLECIGMYEKGIIFNNDPALWKAVRTYFMKALSGPGLVR
 8 KYGDIVRVWINGEETLILSRASAVHHVLKNRKYTSRFGSKQGLSCIGMNEKGIIFNNNVALWKKIRTYFTKALTGPNLQQ
 9 MYGEFMRVWIGGEETLIISKSSSIFHIMKHNHYTCRFGSKLGLECIGMHEKGIMFNNNPALWKAVRPFFTKALSGPGLVR
10 MYGEFIRVWICGEETLIISKSSSMFHVMKHSHYVSRFGSKPGLQCIGMHENGIIFNNNPALWKVVRPFFMKALTGPGLVQ
11 KYGDMVRVWISGEETLVLSRPSAVYHVLKHSQYTSRFGSKLGLQCIGMHEQGIIFNSNVTLWRKVRTYFAKALTGPGLQR
12 TYGDFVRVWISGEETFIISKSSSVSHVMKHWHYVSRFGSKLGLQCIGMYENGIIFNNNPAHWKEIRPFFTKALSGPGLVR
13 KYGDIVRVWINGEETLILSRSSAVHHVLKNGNYTSRFGSIQGLSYLGMNERGIIFNNNVTLWKKIRTYFAKALTGPNLQQ
14 MYGEFMRVWVCGEETLIISKSSSMFHVMKHSHYISRFGSKLGLQFIGMHEKGIIFNNNPALWKAVRPFFTKALSGPGLVR
15 VYGEFMRVWISGEETLIISKSSSMFHIMKHNHYSSRFGSKLGLQCIGMHEKGIIFNNNPELWKTTRPFFMKALSGPGLVR
16 TYGEFVRVWISGEETFIISKSSSVFHVMKHWNYVSRFGSKLGLQCIGMYENGIIFNNNPAHWKEIRPFFTKALSGPGLVR
17 MYGEFMRVWISGEETLIISKSSSMVHVMKHSNYISRFGSKRGLQCIGMHENGIIFNNNPSLWRTVRPFFMKALTGPGLIR

         170       180       190       200       210       220       230       240
           |         |         |         |         |         |         |         |
 1 MVTVCADSITKHLDKLEEVRNDLGYVDVLTLMRRIMLDTSNNLFLGIPLDEKAIVCKIQGYFDAWQALLLKPDIFFKIP-
 2 MVEVCVESIKQHLDRLGEVTDTSGYVDVLTLMRHIMLDTSNMLFLGIPLDESAIVKKIQGYFNAWQALLIKPNIFFKIS-
 3 MVTVCVESVNNHLDRLDEVTNALGHVNVLTLMRRTMLDASNTLFLRIPLDEKNIVLKIQGYFDAWQALLIKPNIFFKIS-
 4 TLEICITSTNTHLDNLSHLMDARGQVDILNLLRCIVVDISNRLFLGVPLNEHDLLQKIHKYFDTWQTVLIKPDVYFRLAW
 5 SVDVCVSATNKQLNVLQEFTDHSGHVDVLNLLRCIVVDVSNRLFLRIPLNEKDLLIKIHRYFSTWQAVLIQPDVFFRLN-
 6 TMEICTTSTNSHLDDLSQLTDAQGQLDILNLLRCIVVDVSNRLFLGVPLNEHDLLQKIHKYFDTWQTVLIKPDVYFRLD-
 7 MVTVCADSITKHLDKLEEVRNDLGYVDVLTLMRRIMLDTSNNLFLGIPLDEKAIVCKIQGYFDAWQALLLKPEFFFKFS-
 8 TVEVCVTSTQTHLDNLSSL----SYVDVLGFLRCTVVDISNRLFLGVPVDEKELLQKIHKYFDTWQTVLIKPDIYFKFS-
 9 MVTVCADSITKHLDKLEEVRNDLGYVDVLTLMRRIMLDTSNNLFLGIPLDESALVHKVQGYFDAWQALLLKPDIFFKIS-
10 MVAICVGSIGRHLDKLEEVTTRSGCVDVLTLMRRIMLDTSNTLFLGIPMDESAIVVKIQGYFDAWQALLLKPNIFFKIS-
11 TLEICTMSTNTHLDGLSRLTDAQGHVDVLNLLRCIVVDISNRLFLDVPLNEQNLLFKIHRYFETWQTVLIKPDFYFRLK-
12 MIAICVESTTEHLDRLQEVTTELGNINALNLMRRIMLDTSNKLFLGVPLDENAIVLKIQNYFDAWQALLLKPDIFFKIS-
13 TVDVCVSSIQAHLDHLDSL----GHVDVLNLLRCTVLDISNRLFLNVPLNEKELMLKIQKYFHTWQDVLIKPDIYFKFR-
14 MVTICADSITKHLDRLEEVCNDLGYVDVLTLMRRIMLDTSNMLFLGIPLDESAIVVNIQGYFDAWQALLLKPDIFFKIS-
15 MVTVCAESLKTHLDRLEEVTNESGYVDVLTLLRRVMLDTSNTLFLRIPLDESAIVVKIQGYFDAWQALLIKPDIFFKIS-
16 MIAICVESTIVHLDKLEEVTTEVGNVNVLNLMRRIMLDTSNKLFLGVPLDESAIVLKIQNYFDAWQALLLKPDIFFKIS-
17 MVEVCVESIKQHLDRLGDVTDNSGYVDVVTLMRHIMLDTSNTLFLGIPLDESSIVKKIQGYFNAWQALLIKPNIFFKIS-
```

```
           250       260       270       280       290       300       310       320
            |         |         |         |         |         |         |         |
 1 WLYRKYEKSVKDLKEDMEILIEKKRRRIFTAEKLEDCMDFATELILAEKRGELTKENVNQCILEMLIAAPDTMSVTVFFM
 2 WLYRKYERSVKDLKDEIAVLVEKKRHKVSTAEKLEDCMDFATDLIFAERRGDLTKENVNQCILEMLIAAPDTMSVTLYFM
 3 WLSRKHQKSIKELRDAVGILAEEKRHRIFTAEKLEDHVDFATDLILAEKRGELTKENVNQCILEMMIAAPDTLSVTVFFM
 4 WLHGKHKRDAQELQDAIAALIEQKRVQLTRAEKFDQ-LDFTGELIFAQSHGELSTENVRQCVLEMIIAAPDTLSISLFFM
 5 FVYKKYHLAAKELQDEMGKLVEQKRQAINNMEKLDE-TDFATELIFAQNHDELSVDDVRQCVLEMVIAAPDTLSISLFFM
 6 WLHRKHKRDAQELQDAITALIEQKKVQLAHAEKLDH-LDFTAELIFAQSHGELSAENVRQCVLEMVIAAPDTLSISLFFM
 7 WLYKKHKESVKDLKENMEILIEKKRCSIITAEKLEDCMDFATELILAEKRGELTKENVNQCILEMLIAAPDTLSVTVFFM
 8 WIHQRHKTAAQELQDAIESLVERKRKEMEQAEKLDN-INFTAELIFAQGHGELSAENVRQCVLEMVIAAPDTLSISLFFM
 9 WLYRKYEKSVKDLKDAMEILIEEKRHRISTAEKLEDSMDFTTQLIFAEKRGELTKENVNQCVLEMMIAAPDTMSITVFFM
10 WLYKKYEKSVKDLKDAIDILVEKKRRRISTAEKLEDHMDFATNLIFAEKRGDLTRENVNQCVLEMLIAAPDTMSVSVFFM
11 WLHDKHRNAAQELHDAIEDLIEQKRTELQQAEKLDN-LNFTEELIFAQSHGELTAENVRQCVLEMVIAAPDTLSISVFFM
12 WLCKKYKDAVKDLKGAMEILIEQKRQKLSTVEKLDEHMDFASQLIFAQNRGDLTAENVNQCVLEMMIAAPDTLSVTLFFM
13 WIHHRHKTATQELQDAIKRLVDQKRKNMEQADKLDN-INFTAELIFAQNHGELSAENVTQCVLEMVIAAPDTLSLSLFFM
14 WLCRKYEKSVKDLKDAMEILIAEKRHRISTAEKLEDSIDFATELIFAEKRGELTRENVNQCILEMLIAAPDTMSVSVFFM
15 WLYKKYEKSVKDLKDAIEVLIAEKRRRISTEEKLEECMDFATELILAEKRGDLTRENVNQCILEMLIAAPDTMSVSLFFM
16 WLCKKYEEAAKDLKGAMEILIEQKRQKLSTVEKLDEHMDFASQLIFAQNRGDLTAENVNQCVLEMMIAAPDTLSVTLFIM
17 WLYRKYERSVKDLKDEIEILVEKKRQKVSSAEKLEDCMDFATDLIFAERRGDLTKENVNQCILEMLIAAPDTMSVTLYVM

           330       340       350       360       370       380       390       400
            |         |         |         |         |         |         |         |
 1 LFLIAKHPQVEEELMKEIQTVVGERDIRNDDMQKLEVVENFIYESMRYQPVVDLVMRKALEDDVIDGYPVKKGTNIILNI
 2 LLLVAEYPEVEAAILKEIHTVVGDRDIKIEDIQNLKVVENFINESMRYQPVVDLVMRRALEDDVIDGYPVKKGTNIILNI
 3 LCLIAQHPKVEEALMKEIQTVLGERDLKNDDMQKLKVMENFINESMRYQPVVDIVMRKALEDDVIDGYPVKKGTNIILNI
 4 LLLLKQNPDVELKILQEMNAVLAGRSLQHSHLSGLHILESFINESLRFHPVVDFTMRRALDDDVIEGYEVKKGTNIILNV
 5 LLLLKQNSVVEEQIVQEIQSQIGERDVESADLQKLNVLERFIKESLRFHPVVDFIMRRALEDDEIDGYRVAKGTNLILNI
 6 LLLLKQNPDVELKILQEMDSVLAGQSLQHSHLSKLQILESFINESLRFHPVVDFTMRRALDDDVIEGYNVKKGTNIILNV
 7 LFLIAKHPQVEEAIVKEIQTVIGERDIRNDDMQKLKVVENFIYESMRYQPVVDLVMRKALEDDVIDGYPVKKGTNIILNI
 8 LLLLKQNPHVELQLLQEIDTIVGDSQLQNQDLQKLQVLESFINECLRFHPVVDFTMRRALFDDIIDGHRVQKGTNIILNT
 9 LFLIANHPQVEEELMKEIYTVVGERDIRNDDMQKLKVVENFIYESMRYQPVVDFVMRKALEDDVIDGYPVKKGTNIILNI
10 LFLIAKHPSVEEAIMEEIQTVVGERDIRIDDIQKLKVVENFIYESMRYQPVVDLVMRKALEDDVIDGYPVKKGTNIILNI
11 LLLLKQNAEVERRILTEIHTVLGDTELQHSHLSQLHVLECFINEALRFHPVVDFSYRRALDDDVIEGFRVPRGTNIILNV
12 LILIAEHPTVEEEMMREIETVVGDRDIQSDDMPNLKIVENFIYESMRYQPVVDLIMRKALQDDVIDGYPVKKGTNIILNI
13 LLLLKQNPHVEPQLLQEIDAVVGERQLQNQDLHKLQVMESFIYECLSFHPVVDFTMRRALSDDIIEGYRISKGTNIILNT
14 LFLIAKHPQVEEAIIREIQTVVGERDIRIDDMQKLKVVENFINESMRYQPVVDLVMRKALEDDVIDGYPVKKGTNIILNL
15 LFLIAKHPNVEEAIIKEIQTVIGERDIKIDDIQKLKVMENFIYESMRYQPVVDLVMRKALEDDVIDGYPVKKGTNIILNI
16 LILIADDPTVEEKMMREIETVMGDREVQSDDMPNLKIVENFIYESMRYQPVVDLIMRKALQDDVIDGYPVKKGTNIILNI
17 LLLIAEYPEVETAILKEIHTVVGDRDIRIGDVQNLKVVENFINESLRYQPVVDLVMRRALEDDVIDGYPVKKGTNIILNI

           410       420       430       440       450       460       470       480
            |         |         |         |         |         |         |         |
 1 GRMHRLEFFPKPNEFTLENFAKNVPYR-YFQPFGFGPRACAGKYIAMVMMKVTLVILLRRFQVQTPQDRCVEKMQKKNDL
 2 GRMHRLEYFPKPNEFTLENFEKNVPYR-YFQPFGFGPRGCAGKYIAMVMMKVVLVTLLRRFQVKTLQKRCIENIPKKNDL
 3 GRMHKLEFFPKPNEFTLENFEKNVPYR-YFQPFGFGPRSCAGKFIAMVMMKVMLVSLLRRFHVKTLQGNCLENMQKTNDL
 4 GRMHRSEFFPKPNEFSLDNFQKNVPSR-FFQPFGSGPRSCVGKHIAMVMMKSILVTLLSRFSVCPVKGCTVDSIPQTNDL
 5 GRMHKSEFFQKPNEFNLENFENTVPSR-YFQPFGCGPRACVGKHIAMVMTKAILVTLLSRFTVCPRHGCTVSTIKQTNNL
 6 GRMHRSEFFSKPNQFSLDNFHKNVPSR-FFQPFGSGPRSCVGKHIAMVMMKSILVALLSRFSVCPMKACTVENIPQTNNL
 7 GRMHRLEFFPKPNEFTLENFAKNVPYR-YFQPFGFGPRACAGKYIAMVMMKVTLVILLRRFQVQTPQDRCVEKMQKKNDL
 8 GRMHRTEFFHKANEFSLENFQKNTPRR-YFQPFGSGPRACVGRHIAMVMMKSILVTLLSQYSVCPHEGLTLDCLPQTNNL
 9 GRMHRLEFFPKPNEFTLENFAKNVPYR-YFQPFGFGPRACAGKYIAMVMMKVILVTLLRRFQVQTQQGQCVEKMQKKNDL
10 GRMHRLEFFPKPNEFTLENFAKNVPYR-YFQPFGFGPRGCAGKYIAMVMMKVILVTLLRRFQVKALQGRSVENIQKKNDL
11 GRMHRSEFYPKPADFSLDNFNKPVPSR-FFQPFGSGPRSCVGKHIAMVMMKAVLLMVLSRFSVCPEESCTVENIAHTNDL
12 GRMHKLEFFPKPNEFSLENFEKNVPSR-YFQPFGFGPRSCVGKFIAMVMMKAILVTLLRRCRVQTMKGRGLNNIQKNNDL
13 GRMHRTEFFLKGNQFNLEHFENNVPRPPTFQPFGSGPRACIGKHMAMVMMKSILVTLLSQYSVCTHEGPILDCLPQTNNL
14 GRMHRLEFFPKPNEFTLENFAKNVPYR-YFQPFGFGPRGCAGKYIAMVMMKVVLVTLLRRFHVQTLQGRCVEKMQKKNDL
15 GRMHRLEFFPKPNEFTLENFAKNVPYR-YFQPFGFGPRGCAGKYIAMVMMKAILVTLLRRFHVKTLQGQCVESIQKIHDL
16 GRMHKLEFFPKPNEFSLENFEKNVPSR-YFQPFGFGPRGCVGKFIAMVMMKAILVTLLRRCRVQTMKGRGLNNIQKNNDL
17 GRMHRLEYFPKPNEFTLENFEKNVPYR-YFQPFGFGPRSCAGKYIAMVMMKVVLVTLLKRFHVKTLQKRCIENMPKNNDL

           490
            |
 1 SLHPDETSG
```

2 SLHPNEDRH
3 ALHPDESRS
4 SQQPVEEPS
5 SMQPVEEDP
6 SQQPVEEPS
7 SLHPDETSG
8 SQQPVEHHQ
9 SLHPHETSG
10 SLHPDETSD
11 SQQPVEDKH
12 SMHPIERQP
13 SQQPVEHQQ
14 SLHPDETRD
15 SLHPDETKN
16 SMHPIERQP
17 SLHLDEDSP

Drawing 5. An evolutionary tree built from neutral evolutionary distances (NEDs) calculated by assuming a first order approach to equilibrium for codon usage at two fold redundant silent sites. Numbers on branches of the tree correspond to evolutionary time (in million years) estimated from the NEDs using a first order rate constant for pyrimidine-pyrimidine transitions of 3 x 10^{-9} changes per base per year.

Drawing 6. The Notch family, with f_2 values for each of the internal nodes, and K_a/K_s values for each of the branches.

DETAILED DESCRIPTION OF THE INVENTION

This disclosure describes the classes of tools that permit the scientist to generate experimentally testable hypotheses concerning the function of a protein starting from an evolutionary analysis. These are outlined below:

I. Tools that detect change in function within a family of proteins.

A. Ratios of silent to non-silent substitution along specific branches of an evolutionary tree including tools that address normalization issues.

B. Covarion behavior, in which individual residues display different mutability in different branches of a tree.

C. Detecting high absolute rates of amino acid substitution, changes per unit time.

II. Tools that detect conservation of function within a family of proteins.

A. Compensatory changes

B. Homoplasy

C. Absolute conservation within a defined evolutionary distance

III. Tools that identify individual residues involved in changes in functionally significant behavior.

A. Residues changing in episodes with high Ka/Ks values, minus residues changing in episodes with low Ka/Ks values

B. Residues displaying covarion behavior

C. Mapping these residues on to models for the secondary, tertiary, and quaternary structure of proteins.

IV. Tools that identify individual residues involved in conserved of functionally significant behavior

A. Residues suffering compensatory changes

B. Residues displaying homoplasy

C. Mapping these residues on to models for the secondary, tertiary, and quaternary structure of proteins.

V. Tools that involve correlation between the evolutionary histories of two families of proteins

A. Correlating the topology of evolutionary trees in two families of proteins

B. Correlating the connectivity of proteins in a gene family

C. Dating events in the molecular history

D. Correlating evolutionary events in two protein families occuring at approximately the same time

E. Correlating evolutionary events in two protein families that are associated with analogous behavior involving expressed/silent ratios

VI. Tools that involve correlation between the evolutionary history of a family of proteins and the evolutionary history of the organism as known from some source other than genomic sequence data, including paleontology, geology, ecology, ontogeny, phylogeny, or systematics (collectively known as the "non-genomic record".

A. Correlating the topology of an evolutionary trees and the non-genomic record.

B. Correlating features of patterns of evolution in specific branches in the evolutionary tree with the non-genomic record

C. Correlating evolutionary events in several protein families occuring at approximately the same time with the non-genomic record

Many of these tools are new in this disclosure. Others were disclosed in Serial No. 07/857,224 and Serial No. 08/914,375 and are claimed here for the first time. In many cases, elements of novelty and utility can be found by combining these tools. This disclosure will systematically indicate the Applicant's presently preferred combinations, with statements of where the Applicant believes that the state of the prior art requires reference to the priority dates of parent applications, where it does not.

All of the tools have in common the same starting point, a basic evolutionary model based on three parts:

(a) An evolutionary tree that shows the familial relationship between the members of the protein family,

(b) A multiple alignment of the sequences of members of the protein family, which shows the evolutionary relationship between the individual amino acids in the sequences, and

(c) The sequences of ancient proteins that were the ancestors of the contemporary proteins in the family.

Each element of an evolutionary model requires the other two in the reconstruction process. Accordingly, processes for constructing an evolutionary model for a protein family are frequently iterative. These processes are well know in the art, and include parsimony tools [Fit67], maximum likelihood tools [Gon91][Gon96][Tho92], tools for evaluating the probability of an evolutionary model [Gon96], and gamma models [Swo96] [Li97].

Serial No. 08/914,375 disclosed the step-by-step procedure in which the basic evolutionary model for a family of proteins is constructed to support the tools outlined above.

(a) A multiple alignment, an evolutionary tree, and ancestral sequences at nodes in the tree are constructed by methods well known in the art for a set of homologous proteins. These three elements of the description are interlocking, as is well known in the art. The presently preferred methods of constructing ancestral sequences for a given tree is the maximum parsimony methods, as implemented (for example) in the commercially available program MacClade [W. P. Maddison, D. R. Maddison, *MacClade. Analysis of Phylogeny and Character Evolution.* Sinauer Associates, Sunderland MA (1992)]. Alternative methods for reconstructing evolutionary intermediates can now be found with the PAUP program [Swofford, D. L., Olsen, G. J., Waddell, P. J., & Hillis, D. M. (1996) Phylogenetic Inference in *Molecular Systematics* (eds. Hillis, D. M., Moritz, C. & Mable, B. K.) 407-514 (Sinauer Assc., Inc., Sunderland, MA, 1996)] and using the maximum likelihood method of the PAML program [Yang, Z. H. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13,** 555-556 (1997)]. Trees are compared based on their scores using either maximum parsimony or maximum likelihood criteria, and selected based on considerations of score and correspondence to known facts. Step (a) is part of the process used to generate the predictions of secondary structure using the method disclosed in Serial No. 07/857,224.

(b) A corresponding multiple alignment is constructed by methods well known in the art for the DNA sequences that encode the proteins in the protein family. The multiple alignment is constructed in parallel with the protein alignment. In regions of gaps or ambiguities, the amino acid sequence alignment can be adjusted to give the alignment with the most parsimonious DNA tree. The presently preferred method of constructing ancestral DNA sequences for a given tree is the maximum parsimony method. The DNA and protein trees and multiple alignments must be congruent, meaning that when amino acids are aligned in the protein alignment, the corresponding codons are aligned in the DNA alignment. Likewise, the connectivity of the two evolutionary trees must show the same evolutionary relationships. In regions where the connectivity of the amino acid tree is not uniquely defined by the amino acid sequences, the tree that gives the most parsimonious DNA tree is used to decide between two trees or reconstructions of equal value. Finally, the ancestral amino acids reconstructed at nodes in the tree must correspond to the reconstructed codons at those nodes. When the ancestral sequences are ambiguous, and where the DNA sequences cannot resolve the ambiguity, the reconstructed DNA sequences must be ambiguous in parallel. Approximate reconstructions are valuable even when exact reconstructions are not possible from available data, and the tree is preferably constrained to correspond to evolutionary relationships between proteins inferred from biological data (e.g., cladistics).

(c) Mutations in the DNA sequences are then assigned to each branch of the DNA evolutionary tree. These may be fractional mutations to reflect ambiguities in the sequences at the nodes of the tree. When ambiguities are encountered, alternatives are weighted equally. Mutations along each branch are then assigned as being "silent", meaning that they do not have an impact on the encoded protein sequence, and "expressed", meaning that they do have an impact on the encoded protein sequence. Fractional assignments are made in the case of ambiguities in the reconstructed sequences at nodes in a tree.

As disclosed in Serial No 08/914,375, the quality of a multiple alignment and the precision of the reconstructed ancestral sequences decreases if proteins are included in the family with sequences diverging by over 150 PAM units, where a PAM unit is the number of point accepted mutations per 100 amino acids. For this reason, families are most preferably constructed with a tree "width" (the distance between the two most divergent proteins in the family) of 150 PAM units or less. Some variation is, of course, desired. Therefore, the PAM width of the tree is preferably more than 50 PAM units. Also referred are well articulated trees. In principle, the more sequences in the tree, the more valuable an evolutionary analysis of the tree becomes.

With the emergence of massive amounts of sequence information as a result of genome projects, the ability to construct detailed evolutionary histories of protein families will increase. This will make the inventions disclosed herein of still greater value, as is appreciated by one of ordinary skill in the art.

One key inventive feature of Serial No 07/857,224 was that an evolutionary analysis had additional value when placed within well defined. One key inventive feature of Serial No 08/914,375 was that an evolutionary analysis gained additional value when it involved analysis of explicitly reconstructed intermediates in the evolutionary tree. These inventive concepts are at the core of all of the tools outlined above.

Another key inventive feature of Serial No 08/914,375 was that an evolutionary analysis gained additional value when it is correlated with the non-genomic record. This inventive concepts is at the core of all of the tools in class VI outlined above.

Another key inventive feature of Serial No 08/914,375 involved the use of a natural organization to generate a rapidly searchable database. As disclosed in the specification to Serial No 08/914,375, when *all* of the genomes of all of the organisms on planet Earth are completed, all protein sequences will be easily recognizable as members of one of ca. 10,000-100,000 *nuclear families*, protein sequence modules 50-500 amino acids long that are related by common ancestry. This conclusion reflects the well known fact that all organisms on the planet are descendants of a single ancestor. In the course of producing the diversity of organisms now on Earth, divergent evolution also produced the diversity of molecular genetic sequences within nuclear families.

As disclosed in the specification to Serial No 08/914,375, this permits a naturally organized database. The ancestral sequences and the predicted secondary structures associated with the families are surrogates for the sequences and structures of the individual proteins that are members of the family. The reconstructed ancestral sequence represents in a single sequence all of the sequences of the descendent proteins. The predicted secondary structure associated with the ancestral sequence represents in a single structural model all of the core secondary structural elements of the descendent proteins. Thus, the ancestral sequences can replace the descendent sequences, and the corresponding core secondary structural models can replace the secondary structures of the descendent proteins.

This makes it possible to define two surrogate databases, one for the sequences, the other for secondary structures. The first surrogate database is the database that collects from each of the families of proteins in the databases a single ancestral sequence, at the point in the tree that most accurately approximates the root of the tree. If the root cannot be determined, the ancestral sequence chosen for the surrogate sequence database is near the center of mass of the tree. The second surrogate database is a database of the corresponding secondary structural elements. The surrogate databases are much smaller than the complete databases that contain the actual sequences or actual structures for each protein in the family, as each ancestral sequence represents many descendent proteins. Further, because there is a limited number of protein families on the planet, there is a limit to the size of the surrogate databases. Based on our work with partial sequence databases [Gonnet et al., op. cit. 1992], we expect there to be fewer than 10,000 families as defined by steps (a) through (e).

Searching the surrogate databases of the instant invention for homologs of a probe sequence thus proceeds in two steps. In the first, the probe sequence (or structure) is matched against the database of surrogate sequences (or structures). As there will be on the order of 10000 families of proteins as defined by steps (a) through (e) after all the genomes are sequenced for all of the organisms on earth, there will be only on the order of 10000 surrogate sequences to search. Thus, this search will be far more rapid than with the complete databases. A probe protein sequence (or DNA sequence in translated form) can be exhaustively matched [Gonnet et al., op. cit. 1992] against this surrogate database (that is, every subsequence of the probe sequence will be matched against every subsequence in the ancestral proteins) more rapidly than it could be matched against the complete database.

Should the search yield a significant match, the probe sequence is identified as a member of one of the families already defined. The probe sequence is then matched with the members of this family to determine where it fits within the evolutionary tree defined by the family. The multiple alignment, evolutionary tree, predicted secondary structure and reconstructed ancestral sequences may be different once the new probe sequence is incorporated into the family. If so, the different multiple alignment, evolutionary tree, and predicted secondary structure are recorded, and the modified reconstructed ancestral sequence and structure are incorporated into their respective surrogate databases for future use.

The advantage of this data structure over those presently used is apparent. As presently organized, sequence and structure databases treat each entry as a distinct sequence. Each new sequence that is determined increases the size of the database that must be searched. The database will grow roughly linearly with the number of organismal genomes whose sequences are completed, and become increasingly more expensive to search.

The surrogate database will not grow linearly. Most of the sequence families are already represented in the existing database. Addition of more sequences will therefore, in most cases, simply refine the ancestral sequences and associated structures. In any case, the total number of sequences and structures in their respective databases will not grow past ca. 10000, the estimate for the total number of sequence families that will be identifiable after the genomes of all organisms on earth are sequenced. If a dramatically new class of organism is identified, this estimate may grow, but not exponentially (as is the growth of the present database).

Since Serial No. 08/914,375 was filed, other databases have emerged that offer some precomputed families. Most noteworthy are Pfam [Bat00] and ProDom [Cor00].

Serial No 07/857,224 disclosed methods to identify residues, secondary structural elements, and evolutionary episodes that are involved in functional adaptation

Further, during episodes of rapid sequence evolution, amino acid substitutions will be concentrated in secondary structural elements defined by the method claimed in Serial No. 07/857,224. These are secondary structural elements that are important in the acquisition of new function. A general method for identifying secondary structural elements that contribute to the origin of new biological function is comprised of identifying an element in the predicted secondary structure model where the corresponding section of the gene has a high ratio of expressed to silent changes.

4. Identification of *in vitro* behaviors that contribute to physiological function.

In vitro experiments in biological chemistry extract data on proteins and nucleic acids (for example) that are removed from their native environment, often in pure or purified states. While isolation and purification of molecules and molecular aggregates from biological systems is an essential part of contemporary biological research, the fact that the data are obtained in a non-native environment raises questions concerning their physiological relevance. Properties of biological systems determined *in vitro* need not correspond to those *in vivo*, and properties determined *in vitro* need have no biological relevance *in vivo.*

To date, there has been no simple way to say whether or not biological behaviors are important physiologically to a host organism. Even in those cases where a relatively strong case can be made for physiological relevance (for example, for enzymes that catalyze steps in primary metabolism), it has proven to be difficult to decide whether individual properties of that enzymes (k_{cat}, K_m, kinetic order, stereospecificity, etc.) have physiological relevance. Especially difficult, however, is to ascertain which behaviors measures *in vitro* play roles in "higher" function in metazoa, including development, regulation, reproduction, digestion.

A general method to determine whether a behavior measured *in vitro* is important to the evolution of new physiological function is comprised of the following steps:

(a) Prepare in the laboratory proteins that have the reconstructed sequences corresponding to the ancestral proteins before, during, and after the evolution of new biological function, as revealed by an episode of high expressed to silent ratio of substitution in a protein. This high ratio compels the conclusion that the protein itself serves a physiological role.

(b) Measure in the laboratory the behavior in question in ancestral proteins before, during, and after the evolution of new biological function, as revealed by an episode of high expressed to silent ratio of substitution. Those behaviors that increase during this episode are deduced to be important for physiological function. Those that do not are not.

We now discuss using the basic evolutionary model in the context of tools that generate hypotheses concerning function within and between protein families.

I. Tools that detect change in function within a family of proteins.

A. Ratios of silent to non-silent substitution along specific branches of an evolutionary tree including tools that address normalization issues.

As discussed in Serial No. 07/857,224, during the divergent evolution of two proteins from a common ancestor, mutations of two types accumulate. The first have no impact on the ability of the host organism to survive, select a mate, and reproduce; these are called "neutral" mutations. The second influence the behavior of the protein in a way that influences the ability of the organism to survive, select a mate, and reproduce. These are termed "adaptive mutations." When evolving a new function, proteins undergo an episode of rapid sequence evolution that corresponds to adaptive "positive selection", as is well known in the art [Kreitman, M., Akashi, H. Ann. Rev. Ecol. Syst. 26, 403-422 (1995)].

Given a basic evolutionary model for a protein family, we can begin to search for sequence details that are indicative of function. For example, the genetic code is degenerate. Some mutations randomly introduced into a genome do not alter the encoded amino acid ("silent mutations"). Others do ("non-silent mutations"). When the gene is under no selective pressure at all, it makes no difference to natural selection whether the mutation changes an amino acid or not. Thus, mutations at the level of the gene are (essentially) neutral, and are fixed in a population without regard to whether they are silent or non-silent. The ratio of non-silent to silent changes can be normalized for the number of silent sites in a particular sequence to give K_a and K_s values.

When the function of a protein is constant, non-silent changes are usually detrimental. Non-silent changes are therefore removed by natural selection. Silent changes are not. The K_a/K_s value is therefore lower than unity in a protein divergently evolving under a constant set of functional constraints. Indeed, for many proteins with function that has been established early in natural history (such as cytochromes), the ratio approaches zero. At the start of the evolutionary period where the calculation is done, the protein is already doing its job nearly optimally, and neither needs nor wants to change its amino acids. Conversely, if one reconstructs the evolutionary history of a protein, and identifies an episode in that evolution where the non-silent/silent ratio is very much less than one, the genomic analysis suggests that the protein has a conserved function during that episode.

One of ordinary skill in the art will note that this method assumes that codon selection is not strongly selected in metazoa. This is not true in eubacteria, or in highly expressed genes in yeast, for example. However, there is little evidence in metazoa to suggest that codon usage is strongly selected in multicellular plants and animals (metazoa), including mammals, where most of the ORFs needing analysis for a developmental biology program are studied. Therefore, the presently preferred scope for methods involving the analysis of silent substitutions is in multicellular organisms.

The exact opposite is the case when new function (implying, of course, new behaviors as well) is being engineered into a protein during an episode of evolution. Non-silent changes, those where amino acids are replaced at the level of the protein, are the only way to change the behavior of a protein to perform its new role. Natural selection desires non-silent changes, as these create new behaviors. The K_a/K_S value is high.

The ratio of non-silent to silent changes, normalized for the number of non-silent and silent sites (the K_a/K_S value) was introduced in the 1980s as a way of detecting change in function between proteins at the leaves of trees[Li97]. It was applied to a large number of cases (for an example, see [McD91][Jol89]). Both the Applicant [Tra96] and Stewart and her coworkers [Mes97] extended this method to analyze reconstructed evolutionary events, calculating K_a/K_S values between ancestral nodes in an evolutionary tree, and applied it to individual cases (ribonuclease and lysozyme, respectively). Using this approach, if one reconstructs the evolutionary history of a protein, and identifies an episode in that evolution where the K_a/K_S value is greater than unity, the protein is evolving a new function during that episode.

In practice, K_a/K_S values are not so easily interpretable. Even when the function of a protein is changing, some residues (such as those holding together the fold) cannot change without destroying the ability of the protein to serve as a scaffold for function. Thus, the K_a/K_S value for specific sites can be very high during an episode of divergent evolution, perhaps even much higher than unity. But because K_a/K_S values are calculated for the sequence as a whole, the sites undergoing rapid substitution are counted with "core" sites undergoing slow substitution, giving a K_a/K_S value for the protein as a whole of less than unity.

Likewise, K_a/K_S values are assigned to individual branches of an evolutionary tree. If the evolutionary tree is poorly articulated, a single branch may contain both adaptive and conservative episodes of evolution. In this case, the high K_a/K_S value for the adaptive episode may be diluted by a low K_a/K_S value for the conservative episode. The second problem will, of course, subside as more and more genome sequence projects are completed.

One solution to this problem involves normalization of the K_a/K_S values for a protein family. Here, the average K_a/K_S value for the average branch of the tree is calculated. Thos branches that have a K_a/K_S value an arbitrary factor higher (the presently preferred factor is two fold higher) are then hypothesized to be undergoing a change in function. More preferably, a statistical analysis is performed where the number of sites undergoing changes is determined for each branch length, the average K_a/K_S value is calculated, a statistical model is constructed to assess the distribution of K_a/K_S values on different branches of the tree, and branches that have K_a/K_S values lying more than two standard deviations above the mean are hypothesized to contain a change in function

Serial No. 08/914,375 discussed in greater detail the tools based on the fact that the genetic code is degenerate. More than one triplet codon encodes the same amino acid. Therefore, a mutation in a gene can be either silent (not changing the encoded amino acid) or expressed (changing the encoded amino acid). Especially in multicellular organisms, and most particularly in multicellular animals (metazoa), silent changes are not under selective pressure. In contrast, expressed changes at the DNA level, by changing the structure of the protein that the gene encodes, change the property of the protein.

When examining a protein from higher organisms during a period of evolutionary history where, at the outset of the period, the behavior of a protein is optimized for a specific biological function, and where that function remains constant for the protein throughout the period being examined, changes in the DNA sequence that lead to a change in the sequence of the encoded protein (expressed changes) will diminish the survival value of the protein [Benner, S. A., Ellington, A. D. Interpreting the behavior of enzymes. Purpose or pedigree? *CRC Crit. Rev. Biochem.* **23**, 369-426 (1988)] and therefore will be removed by natural selection. During the same period, silent changes will not be removed by natural selection, but will accumulate at an approximately clock-like rate, as silent changes are approximately neutral, especially in higher organisms. Thus, the ratio of expressed to silent changes will be low during a period of evolution of a protein family where the ancestor and its descendants share a common function.

In contrast, in genes for proteins that are neutrally drifting without functional constraints, the expressed/silent ratio will reflect random introduction of point mutations. Given the genetic code and a typical distribution of amino acid codons within the gene, a ratio of expressed to silent changes will be approximately 2.5 during the period of evolution of a protein family where the ancestor and its descendants have no function.

A third situation concerns a period of evolution where a protein is acquiring a new derived function. The amino acid sequence of the protein at the beginning of this episode will be optimized for the ancestral function, rather than the derived function. Thus, changes in the gene that are expressed in changes in the sequence of the encoded protein that improve the behavior of the protein as is required for the new biological function will be selected for. In proteins in such an evolutionary episode seeking new function, natural selection seeks expressed changes, and the ratio of expressed to silent substitutions at the DNA level will be high during the period of evolution of a protein family where the function of the ancestor has changed with a new function emerging in its descendants. Ratios as high as 4:1 or more are known.

In a family of proteins defined by steps (a) through (e) above, individual periods of evolution are defined by lines between nodes on an evolutionary tree. In step (c), silent and expressed point mutations are assigned to individual periods of evolution. Periods of evolution with high ratios of expressed to silent mutations are episodes where physiological function is rapidly changing. Periods of evolution with low ratios of expressed to silent mutations are episodes where physiological function is slowly changing.

Serial No. 08/914,375 showed the application of this approach applied to the leptin family of proteins. Leptins are present in mice, where they are believed to modulate feeding behavior. Leptin homologs are also present in humans, and the pharmaceutical industry has been excited about exploiting them in the treatment of obesity. The conclusion drawn from this hypothesis is that the leptin protein in humans does not have the same function as the leptin protein in mice.

B. Covarion behavior, in which individual residues display different mutability in different branches of a tree.

Functional changes leave signatures in the patterns of sequence evolution in a protein family. Covarion behavior was detected in alcohol dehydrogenase [Ben89] and superoxide dismutase [Miy95]. As a preliminary study in the past year, elongation factors (EF) serve as an example. These are proteins that have diverged far more slowly; indeed, they are archetypal examples of a protein that performs the "same" function in all three kingdoms of life. In the example, thirty EF-Tu/EF-1α protein sequences were aligned over 380 sites using the alignment program DARWIN. Replacement rates per site for bacterial and eukaryotic EFs were estimated using a gamma-based, maximum likelihood (ML) model for protein sequences (JTT + Γ) and the phylogeny of Baldauf *et al.* [Bal96] for EF-Tu and EF-1α. An α of 0.78 was calculated for the entire tree, with a standard deviation (SD) of 0.05 using parametric bootstrapping (evolutionary simulations) [Swo96]. Interestingly, the α values for the bacterial and eukaryotic subtrees were significantly different from that for the entire tree [0.46 (0.04) and 0.38 (0.04), respectively]. These reductions in α for bacteria and eukaryotes alone are expected of a non-stationary covarion process.

The distribution of rate differences per site between bacterial and eukaryotic EFs is leptokurtotic; i.e., over- and under-represented in the mean and tails versus "shoulders," respectively, relative to the expectations of a normal distribution. Thirty seven percent of the sites have essentially the same rate in the two groups (rate difference of ~0), as expected under a stationary gamma process. However, 18 and 21 sites evolve >2 SD faster in bacteria than eukaryotes, and vice versa, respectively. These 10% of the sites are most responsible for the covarion characteristics of EF-Tu and EF-1α.

Residues displaying abnormal evolutionary behavior were then mapped to a three dimensional model of the protein based on a crystal structure of ET-Tu. These were used to generate structural hypotheses for the different behavioral differences that were known. For example, bacterial EF-Tu binds GDP ~100 fold tighter than GTP. Eukaryotic EF-1α, in contrast, binds both with similar affinities. EF-Tu regenerates its active form by binding to the single-subunit nucleotide exchange factor EF-Ts. EF-1α requires the multi-subunit nucleotide exchange factor EF-1βγδ. EF-1α also interacts with the cytoskeleton and may thereby play a role in cellular transformation and apoptosis. EF-Tu can have no such role in bacteria. Residues were identified that, at the level of hypothesis, are responsible for each of these behavioral differences.

Covarion behavior indicates changing function. It is therefore expected to correlate positively with events with high K_a/K_S ratios. Because K_a/K_S ratios use a silent substitution clock that ticks rapidly, while covarion analysis does not, the two are somewhat complementary.

C. Detecting high absolute rates of amino acid substitution, changes per unit time.

An alternative way to detect changes in function is to measure the number of amino acids substitutions that occur per unit time. This requires that dates be assigned to nodes in an evolutionary tree. This can be done by correlation with the paleontological record, as is well known in the art.

II. Tools that detect conservation of function within a family of proteins.

A. Compensatory changes

The conservation of the overall fold after extensive divergences raises the possibility that amino acid substitutions at one position in a polypeptide chain might be compensated by substitutions elsewhere in a protein. For example, if a Gly at one position inside the folded protein core is replaced by a Trp, it might be necessary to substitute a Trp by a Gly at a position distant in the sequence but near in space to conserve the overall volume of the core, and therefore the overall folded structure. These assume that if a substitution is not compensated, the organism hosting the protein is less fit.

Individual examples of compensatory changes in proteins have been proposed [Oos86], both by analysis of families of natural proteins with known structures [Les80][Les82][Cho82][Alt87][Alts88][Bor90] and in proteins into which point mutations have been introduced by site-directed mutagenesis [Lim89][Lim92][Bal93]. In these examples, amino acid residues distant in the sequence but near in three dimensional space in the folded structure have been observed to undergo simultaneous compensatory variation to conserve overall volume, charge, or hydrophobicity.

Compensatory covariation has been used in the prediction of the tertiary folds. For protein kinase [Ben91], for example, an antiparallel beta sheet was predicted for the core of the first domain because of two specific compensatory changes identified in consecutive strands in the predicted secondary structural model. The subsequently determined crystal structure [Kni91] showed not only that antiparallel beta sheet existed, but that the side chains of the two residues undergoing compensatory covariation were indeed in contact.

Systematic studies have suggested, however, that the compensatory covariation generates only a small signal. The early work by Lesk and Chothia with the globin family found that replacements of hydrophobic residues in the core of the protein fold are usually accommodated by small shifts of secondary structural elements rather than by size complementary amino acid substitutions [Les80][Les82][Cho82]. More recent studies have suggested that a weak compensatory covariation signal might exist [Tay94][Shi94][Göb94][Neh94]. Some authors have doubted, however, that the signal is adequate to be useful in structure prediction [Tay94]. Others have been more optimistic [Neh94][Shi94]. More recently, Chelvanayagam et al. pointed out that the signal might be improved if examples of compensatory covariation were sought within explicit evolutionary context [Che97][Che98].

In the literature, compensatory changes have been sought by comparing the sequences of two extant proteins from contemporary organisms. In principle, any position where an amino acid residue had undergone substitution at any point in the time separating the two proteins via the common ancestor might be paired with any other position that had also suffered substitution in this time. Such an approach is problematic because the evolutionary time separating two contemporary protein sequences can be long; in years, it is twice the time since the most recent common ancestor of the two proteins.

A different way to detect compensatory covariation begins with the recognition that a model for the historical past in a protein family can be inferred from a set of homologous protein sequences These models have three parts: (a) an evolutionary tree, which shows the genealogical relationships between individual proteins in the family, (b) a multiple sequence alignment, which shows the evolutionary relationship between individual nucleotides in the genes encoding each family, and (c) reconstructed sequences of ancestral proteins that are evolutionary intermediates in the tree. Through the reconstruction of ancestral sequences, specific changes in a protein sequence can be assigned to (and isolated to) specific branches of the evolutionary tree. Within the context of a reconstructed model for the historical past, compensatory covariation should appear as two substitutions occurring on the same branch of the evolutionary tree. As these branches can be rather short in length, an analysis based on a reconstructed history of a protein family can identify changes that occur nearly simultaneously. These are expected to be true indicators of compensation. In principle, a weak compensatory covariation signal observed by the comparison of extant sequences should be strengthened by examining individual episodes in divergent evolution as reflected by specific branches in the evolutionary tree.

In preliminary studies, we examined 71 families of proteins from the Master Catalog to learn whether reconstructed ancestral sequences will generate a more useful signal for compensatory covariation than can be obtained by examining extant sequences. We noticed anecdotally that covariation was more likely to occur along branches with low Ka/Ks values. This makes sense, as compensation is necessary only if function is conserved. Case studies developed under this project will test this.

B. Homoplasy

One feature commonly observed in the divergent evolution but not modelled well by even advanced stochastic models is molecular homoplasy, defined as a character similarity that arose independently in different subfamilies of an evolutionary tree [Str00]

Molecular homoplasy is best illustrated by an example (Drawing 3). Homoplasy so defined is the observed phenomenon; no statement is made as to the mechanism by which homoplasy arises. It may reflect selection pressures. The Master Catalog gives us the opportunity to systematically search for molecular homoplasy in the database as a whole.

At one level, homoplasy is simply the statement that selective pressures are forcing the protein to select from a subset of the 20 standard amino acids. Thus, it is similar to the bias that is seen in membrane proteins, for example (where residues are chosen more frequently from a subset of hydrophobic amino acids than in the database as a whole). Homoplasy is more. Not only (in the example) is position 30 limited to A and P, but the selection pressures have toggled between the two more than once in the module's evolutionary history.

This is, of course, a signature that a functional constraint is conserved in the distant branches of the tree protein. For this reason, molecular homoplasy is expected to be a contrarian signature to high K_a/K_s or non-stationary covarion behavior in a protein. We expect it to occur more frequently with proteins that are *not* undergoing functional recruitment.

Some informative features are already evident from preliminary work. For example, a preliminary search of 38 protein families with high resolution crystal structures identified over 2000 examples of molecular homoplasy. These were characterized first by the nature of the amino acids identified. A number of very obvious patterns emerged. First, the majority of the examples involve the interchange of hydrophobic side chains of nearly identical volume. The homoplasy involving I and V was the most frequent. It occurred 230 times in the dataset. The I/V molecular homoplasy was far more abundant than the next most popular hydrophobic/hydrophobic homoplasy, F/Y, which was found 68 times, and the I/L hydrophobic/hydrophobic homoplasy, which was found 44 times. As might be expected, the majority of these were buried in the three dimensional structure of the protein.

In the next phase of work we will ask whether these homoplasies are correlated with homoplasies at other positions in the same sequence in the same branches of the trees. If the functional constraint at the amino acid position are sufficient to permit a protein to confer fitness only if it places one of two residues there, then this constraint might be sufficient to cause compensation, also possibly homoplastic, at a second position nearby in the folded structure of the protein. Further, it is necessary to characterize the branch length (NED or PAM) where the changes occur.

The most interesting homoplasies are those that involve multiple steps. For example, the Pro/Gly homoplasy (at the codon level, CCN to GGN) requires two substitutions. Either of these alone creates a change in the encoded amino acid (CGN, Arg, or GCN, Ala). Observing examples of these without observing the intermediates anywhere else in the tree suggests that selection pressure is remarkably strong at this position, even though two amino acids appear to be nearly equally suited to perform function.

Molecular homoplasy indicates a constraint on structure that implies a constant behavior, which in turn implies a constant function. If this is true, it should correlate negatively with K_a/K_S ratios. That is, homoplasy should be found less frequently in branches separated by a branch with a high K_a/K_S ratio than in branches not separated by such a branch. Case studies developed under this project will develop ways to exploit such a correlation.

C. Absolute conservation within a defined evolutionary distance

As disclosed in Serial No. 07/857,224, residues that are conserved over an entire evolutionary tree are presumed (at the level of hypothesis) to be important for function, especially if they are chosen from the group consisting of Asp, Lys, Arg, Glu, Asn, Cys, His, Gln, Ser, and Thr. As disclosed in that application, however, it is important that the overall PAM width of the tree be considered before constructing hypotheses about the functional role of conserved residues.

III. Tools that identify individual residues involved in changes in functionally significant behavior.

In Serial No. 08/914,375, it was disclosed that during episodes of rapid sequence evolution, amino acid substitutions will be concentrated in secondary structural elements. These are secondary structural elements that are important in the acquisition of new function. These elements might be predicted using the method claimed in Serial No. 07/857,224; they might also be known by X-ray crystallography or n.m.r., for example. As n Serial No. 08/914,375, a general method for identifying secondary structural elements that contribute to the origin of new biological function is comprised of identifying an element in the predicted secondary structure model where the corresponding section of the gene has a high ratio of expressed to silent changes.

In this analysis, we must recognize tthat function involves combinations of behaviors of a protein. Even when function changes, some features of those behaviors are conserved, and this reflects conservation of some features of the sequence as well. In the fumarase/aspartase/adenylosuccinate lyase example discussed above, all three proteins have the same overall fold. For this reason, residues critical to the folding process (for eample, amino acids whose side chains pack tightly into the folded core) will remain conserved even though the overall function of the protein is changing. Relevant to the change in function is, of course, a change in a number of behaviors, for example, the ability to bind a particular small molecule substrate. Residues involved in substrate binding will therefore be changing rapidly during the episode of sequence evolution where function was changing.

The notion that some residues are conserved even when function is chaning is matched by the notion that some residues will be changing even when function is conserved. The latter are those that can drift "neutrally".

Likewise, "function" remains a concept set within Darwinian evolution. That is, a fumarase from a mesophile and a fumarase from a thermophile have analogous function in the sense that they both participate (for example) in the citric acid cycle. However, they have different functions, in that one contributes to fitness in a thermophile (which requires that it have an associated behavior, thermostability) while the other does not. In the epsidoe where the temperature of the environment changes, residues involved in conferring th ermal stability will change, while those involved in determining substrate specificity will not.

Tools that assign, even at the level of hypothesis, which residues are involved in which behavior are extremely valuable. They can be the targets of protein engineering experiments, for example. In these cases, one would like to map residues identified using tools of the instant invention on to a three dimensional structure of a representative member of a protein family.

Already in 1988, the Applicant was using a general form of mapping that showed the utility of this in extracting information about the function of a protein, in this case, alcohol dehydrogenase [Ben88 xxx]. More recently, Lichtarge et al. introduced an evolutionary trace method that defined functionally significant residues as those that are conserved within a family [O. Lichtarge, H. R. Bourne, F. E. Cohen, An evolutionary trace analysis defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342-358 (1996).]. They then used this approach to identify patches on the surface of proteins that contribute to functionality.

As it was published, the evolutionary trace method was related to the method disclosed in Serial No. 07/857,224, and was applied to conserve amino acid residues. The aproach did not contemplate the possibility that function might change within a family of proteins, and the residues important for function would change with it. Indeed, to detect such changes would require tools disclosed in this application and in Serial No. 08/914,375 to be broadly useful.

A. Residues changing in episodes with high Ka/Ks values, minus residues changing in episodes with low Ka/Ks values

We have posited that function is changing during an episode with high K_a/K_s values. As disclosed in Serial No. 08/914,375, individual residues can be identified as changing during that episode, as the basic evolutionary model has sequences reconstructed at each individual node. These are, at the level of hypothesis, residues that are important to functional change.

As one of ordinary skill in the art recognizes, the episode also includes a number of substitutions that have no relevance to function or the change in function, but rather reflect the background, neutral drift. For example, these residues might lie on the surface of the protein, be in contact with bulk solvent, and not have any especially strong functonal constraint that prevents them from diverging. As disclosed in Serial No. 07/857,224, surface residues are likely to be neutrally drifing in many sub-families within an evolutionary tree. For this reason, we can identify residues that are changing along branches of an evolutionary tree that have low K_a/K_S values, and subtract them from residues changing in episodes with high K_a/K_S values. What remains are residues more likely, again at the level of hypothesis, to be involved in the change in function.

Serial No. 07/857,224 disclosed and claimed methods for correlating changes in sequence with changes in the behavior of the protein. This in turn provides a method for identifying behavioral changes that are relevant to the change in function.

B. Residues displaying covarion behavior

Again because the basic evolutionary model includes reconstructed ancestral intermediates, the methods of the instant invention identify specific residues that are displaying covarion behavior. These are residues that are under analogous functional constraints in different sub-families of the tree. This, in turn, implies that these particular residues contribute to a behavior that is conserved for a conserved feature of the function in distant branches of the tree.

C. Mapping these residues on to models for the secondary, tertiary, and quaternary structure of proteins.

Insight into the relationship between function and amino acid sequence can be gained by mapping residues identified by K_a/K_S and covarion analysis onto a three dimensional structure. This identifies, for any particular branch, which residues are involved in changing function. This information is useful when attempting to identify residues that might be changed in a protein engineering experiment, for example.

IV. Tools that identify individual residues involved in conserved of functionally significant behavior

The type of analysis used for class III tools can also be applied to class IV tools.

A. Residues suffering compensatory changes

When a pair of residues suffers compensatory changes during a particular episode of protein sequence evolution, this implies that some physical property of the protein family must be the same at the end of the episode as it was at the beginning. This implies some conserved behavior important across that episode. The episode can, of course, be one where function in some sense is changing. Thus, in the fumarase/aspartase example outlined above, one might identify residues the suffer compensatory changes during episodes where catalytic behavior is changing. These are residues most likely (at the level of hypothesis) to be important for folding, which is conserved over this episode. We can therefore use the methods of the instant invention to identify individual residues involved in conserved of functionally significant behavior

B. Residues displaying homoplasy

Positions that display homoplasy are subject to analogous functional constraints in different branches of the tree. Because of the evolutionary reconstructions in the basic evolutionary model, we know which positions they are are which amino acids involved. Therefore, we use the methods of the instant invention to identify individual residues involved in conserved of functionally significant behavior

C. Mapping these residues on to models for the secondary, tertiary, and quaternary structure of proteins.

Insight into the relationship between function and amino acid sequence can be gained by mapping residues identified by K_a/K_s and covarion analysis onto a three dimensional structure. This identifies, for any particular branch, which residues are involved in changing function. This information is useful when attempting to identify residues that might be changed in a protein engineering experiment, for example.

V. Tools that involve correlation between the evolutionary histories of two families of proteins

Serial No. 07/857,224 introduced in the first useful form the notion of compensatory changes as a way of analyzing divergent evolution in protein sequences. In that application, an example of compensatory covariation was identified that indicated the packing of two beta strands in an antiparallel fashion. A second use for compensatory changes disclosed was as part of a tool to detect disulfide bonds in a protein; cysteines that arise and/or disappear at the same time during the divergent evolution of a protein family frequently form a disulfide bond with each other. Serial No. 08/914,375 extended this notion, noting that the introduction and loss of leptin and the leptin receptor might occur in parallel. The idea behind this analysis is that residues that interact as they contribute to function, subunits that interact as they contribute to function, and even proteins that interact as they contribute to function, display correlated evolution.

Since these applications were filed, various other groups have extended this approach. We review briefly two of the areas where research is active, and make comments on why additional invention is necessary to make these approaches fully useful

A. Correlating the topology of evolutionary trees in two families of proteins

Recently, Pellegrini et al. extended this type of analysis to generate "protein phylogenetic profiles" for different organisms [Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., Yeates, T. O. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles PNAS 96, 4285-4288 1999]. They present a method that assumed that during evolution, proteins that function together tend to be either preserved or eliminated in a new species. They described this property of correlated evolution by characterizing each protein by its phylogenetic profile, a string that encodes the presence or absence of a protein in every known genome. They suggested that proteins having matching or similar profiles strongly tend to be functionally linked. This method of phylogenetic profiling allows us to predict the function of uncharacterized proteins.

More recently, Cohen and his coworkers used phosphoglycerate kinase (PGK), an enzyme that forms its active site between its two domains, to develop a standard for measuring the co-evolution of interacting proteins. The N-terminal and C-terminal domains of PGK form the active site at their interface and are covalently linked. Therefore, they must have co-evolved to preserve enzyme function. By building two phylogenetic trees from multiple sequence alignments of each of the two domains of PGK, they calculated a correlation coefficient for the two trees that quantifies the co-evolution of the two domains. The correlation coefficient for the trees of the two domains of PGK is 0.79, which establishes an upper bound for the co-evolution of a protein domain with its binding partner. Their analysis was extended to ligands and their receptors, using the chemokines as a model [Goh, C.-S., Bogan, A. A., Joachimiak, M., Walther, D., Cohen, F. W. (2000) Co-evolution of Proteins with their Interaction Partners. *J. Mol. Biol.* **xxx**, 283-293.

We have no quarrel with either of these approaches; indeed, they are in some ways covered by the Applicant's earlier disclosures. It should be recognized, however, that these simple approaches that exploit evolutionary analysis are easily defeated by the "ortholog paralog problem", especially when it is coupled with gene loss. Briefly, paralogs are generated when a gene duplication occurs internally within a genome, to create two homologous genes in the same organism.

B. Correlating the connectivity of proteins in a gene family

Eisenberg and his coworkers. Enright [AJ] et al., and others have also suggested that proteins that interact in a pathway might be connected physically in the genome, either as an operon or, in some cases, in a single expressed polypeptide chain. This interesting approach is applicable to only a subset of the database, and is distinct from the tools disclosed here. [Marcotte, E.M., M. Pellegrini, H.L. Ng, D.W. Rice, T.O. Yeates, and D. Eisenberg. 1999. Detecting protein function and protein-protein interactions from genome sequences. Science. 285: 751-753]

C. Dating events in the molecular history

A key element to using evolutionary analysis of correlated change in protein families is to establish that the changes being interpreted as evidnce that two proteins interact as they function is to show that the changes are contemporaneous, that is, they occur near the same time. This requires tools that date, if only approximately, events in the molecular evolutionary tree using sequence data.

Early hope that protein sequences might change in a "clock-like" fashion [Can82], with a small number of rate constants describing the rate of change at most positions in most proteins in most organisms, has given way to the reality that the evolution of protein sequences is marked by episodes of rapid and slow evolution [Mes97]. These correspond to changing and conserved function within the protein family, arising in turn from adaptive and purifying natural selection, respectively. This makes methods based on protein sequence divergence unreliable for dating the divergence of protein sequences.

One well known approach to avoid (to a large extent, at least in metzoans) the influence of purifying and adaptive selection on the interpretation of molecular history is to examine changes in non-coding regions of DNA [Li97]. These include introns and substitutions, generally at the third position of a codon, that do not change the encoded amino acid. These arise because the genetic code is redundant for many amino acids. This approach assumes that silent substitutions at the DNA level have little or no impact on fitness (are neutral or nearly neutral) at the level of the organism. While this is almost certainly *not* a good approximation in microorganisms, the approximation appears to be serviceable for metazoans (multicellular animals) and plants, presumably because macrophysiology is more visible to selective forces than genome sequence itself in multicellular organisms.

Even silent substitutions are problematic as a molecular clock, however. From a chemical perspective, interconverting the four standard nucleobases A, G, T, and G involves 12 rate constants that need not be identical [Nei86]. Some models distinguish between transitions (purines replaced by purines, or pyrimidines replaced by pyrimidines) and transversions (purines replaced by pyrimidines, or pyrimidines replaced by purines), but otherwise group the rate processes together. This problem is revisited frequently in the literature [Nei86]. The most widely used method was developed by Li [Li85] with modifications by Pamilo and Bianchi [Pam93]. This method aggregates four fold redundant and two fold redundant sites, analyzes nucleotide substitution at positions where the encoded amino acid has not changed at the same time as it analyzes substitution at positions where the encoded amino acid has changed, and adopts a classification of different types of substitutions based on physical chemical characteristics of amino acids.

Disclosed here for the first time, the Applicant has discovered good part of the inconsistency in the dating generated by these methods can be eliminated if one focuses on relatively homogeneous chemical processes. In particular, transitions accumulate over large periods of (for example) vertebrate history with remarkable constancy, with a pseudo first order rate constant of 3.0×10^{-9} changes/base/year. A tool based on this discovery begins by extracting aligned pairs of codons from a pairwise alignment where two fold redundant amino acids (CDEFHKNQY) are conserved. Substitution at the silent position is then modelled using an exponential "approach to equilibrium" rate law, where f2 is the fraction of the codons encoding conserved 2FR amino acids that are themselves conserved: $f2 = [0.5 \bullet \exp(-kt)] + 0.5$, where k is a single pseudo first order rate constant for transitions, and t is the time. The neutral evolutionary distance (NED) between two genes x and y is defined by $\mathrm{NED}_{x,y} = kt_{x,y} = -\ln[(f2_{x,y}+0.5)/0.5]$.

NEDs represent one choice in a trade-off, between the instinct of a statistician (to maximize the number of characters being examined, and hence minimize error due to fluctuation) and the instinct of an organic chemist (to seek homogeneous rate processes, and hence minimize systematic error due to aggregation of different kinds of events).

The NED is a measure of evolutionary distance, not evolutionary time. If one knows the rate constant, and assumes that k is constant over the period of evolutionary history being examined, one can calculate the time of divergence. Given the same assumption and the date of evolutionary divergence of two sequences, one can calculate k. As distances, NEDs are additive, should obey the triangle inequality, and display other features that permit them to be used to build evolutionary trees.

The transition-based two fold NED turned out to be remarkably robust measures of evolutionary time. When calibrated using datable fossil divergences back to the divergence of fish from land vertebrates, a single lineage rate constant of 3 x 10^{-9} changes per base per year was obtained in many of the cases we examined, applicable (within error) to the divergence of fish from mammals, reptiles and birds from mammals, primates from artiodactyls, and artiodactyl genera from other artiodactyl genera. NEDs built from four fold redundant systems were far less consistent.

One of the key issues in the development of evolutionary models is assigning ranges of geological dates to nodes in the tree. Early hope that protein sequences might change in a "clock-like" fashion, with a small number of rate constants describing the rate of most amino acid substitutions in most proteins in most organisms, has given way to the reality that the evolution of protein sequences is marked by episodes of rapid and slow evolution. These correspond to changing and conserved function within the protein family, arising from adaptive and purifying natural selection respectively. This makes protein sequence similarity (for example, point accepted mutations per 100 amino acids, or PAM units) unreliable for dating the divergence of protein sequences.

One well known approach to avoid the influence of purifying and adaptive selection on the interpretation of molecular history is to examine changes in non-coding regions of DNA. These include introns and substitutions, generally at the third position of a codon, that do not change the encoded amino acid. These arise because the genetic code is redundant for many amino acids. Amino acids encoded by four synonymous codons (A_4's) are valine, alanine, threonine, proline and glycine. Amino acids encoded by two synonymous codons (A_2's) are cysteine, aspartic acid, glutamic acid, phenylalanine, histidine, lysine, asparagine, glutamine, and tyrosine. One amino acid (isoleucine) is encoded by three synonymous codons (A_3's). These patterns are found in the eukaryotic nuclear code; other codes exist, of course.

This approach has a chance of working if silent substitutions at the DNA level have little or no impact on fitness at the level of the organism. While this is almost certainly not a good approximation in microorganisms (at least for some codons in highly expressed genes), the approximation appears to be serviceable for metazoans (multicellular animals), presumably because redundant codon exchange does not change the structure or the behavior of any functioning protein, and the structure and behavior of functioning proteins, together with the consequent macrophysiology, is more visible to selective forces than genome sequence itself. The approach is now empirically shown to be reliable within chordates.

Even silent substitutions are problematic as a molecular clock, however. From a chemical perspective, interconversion of the four standard nucleobases A, G, T, and G involves 12 rate constants that need not be identical (there is a large literature on this; see for example [Nei86]). Simpler models have distinguish between transitions (purines replaced by purines, or pyrimidines replaced by pyrimidines) and transversions (purines replaced by pyrimidines, or pyrimidines replaced by purines), but otherwise grouped the rate processes together.

This problem has been revisited frequently in the literature. The most widely used method (indeed, the one implemented in the present version of the Master Catalog when assigning K_a/K_s values, following some adaptations that we made, Schreiber, Benner unpublished) was developed by Li [Li85] with modifications by Pamilo and Bianchi [Pam93] following a suggestion by Kimura.

In the previous funding period, we developed and tested a NEDs as a tool for dating sequence divergences Table 1). NEDs turned out to be remarkably robust measures of evolutionary time. When calibrated using datable fossil divergences back to the divergence of fish from land vertebrates, a single lineage rate constant of 3 x 10^{-9} changes per base per year was obtained in many of the cases we examined, applicable (within error) to the divergence of fish from mammals, reptiles and birds from mammals, primates from artiodactyls, and artiodactyl genera from other artiodactyl genera. Statistical analysis suggests that >80% of the variance arises from simple statistical fluctuation. This suggests the absence of "hot spots" and other non-stochastic variation at the 2-fold degenerate sites in the genome. Again, relatively expensive tools (such as full blown ML tools) gave insignificantly different results than relatively cheap tools (such as the Pamilio-Bianchi approach) in a series of test cased that were applied in parallel.

Table. Average NED values for Pairs of Proteins Extracted from Humans, Pigs, Oxen, Rabbit, Rat, and Mouse

Species 1	Species 2	Number of pairs	kt (range) (NED)	Date (fossil) MYA	*k* (calc.) x 10^9 changes/base/year	*k* (average) x 10^9
Human	Pig	225	0.3990	80	2.5	
Human	Ox	410	0.3800	80	2.4	2.4
Pig	Ox	140	0.2755	60	2.3	
Rabbit	Human	203	0.4845	80	3.0	
Rat	Human	584	0.4893	80	3.0	3.1
Mouse	Ox	147	0.5130	80	3.2	
Mouse	Human	918	0.4988	80	3.1	
Mouse	Rabbit	87	0.5083	60	4.2	5.2
Mouse	Rat	926	0.2470	20	6.2	

D. Correlating evolutionary events in two protein families occuring at approximately the same time

Given approximate dates, we can now provide a more useful tool to correlate events occurring in two trees. A duplication in family 1 that is occurring near the time as a duplication occurring in family 2 is hypothesized to indicate that the two families (and, in particular, the proteins arising from the duplication) interact when they function. Conversely, and frequently quite usefully, a duplication in family 1 that did *not* occur near the time as a duplication occurring in family 2 is hypothesized to indicate that the two proteins arising from the duplication do *not* interact when they function. These hypotheses are ueful when designing two-hybrid systems, for example, to detect protein-protein contacts.

E. Correlating evolutionary events in two protein families that are associated with analogous behavior involving expressed/silent ratios

When there is a duplication, the question arises: Which of the derived genes is performing the derived function, and which is performing the ancestral function? According to the method of this invention, the derived protein is the one connected to the node where the duplication has occurred via the higher K_a/K_s value. This concept supports a useful tool to correlate events occurring in two trees. A duplication in family 1 that is occurring near the time as a duplication occurring in family 2 is hypothesized to indicate that the proteins arising from the duplication from the branch having the higher K_a/K_s value in one tree interact when they function with the proteins arising from the duplication from the branch having the higher K_a/K_s value in one tree interact when they function with the. Conversely, and frequently quite usefully, when examining two contemporarneous duplication events in two separate families, the proteins in family 1 that do *not* interact with the proteins in family 2 are those that are not joined to their respective nodes via branches that display, during contemporaneous periods of evolution, high K_a/K_s values.

As one of ordinary skill in the art will appreciate, this approach is quite general, and can be applied with covarion behavior, compensatory substitution, homoplasy, and even levels of high sequence conservation.

VI. Tools that involve correlation between the evolutionary history of a family of proteins and the evolutionary history of the organism as known from some source other than genomic sequence data, including paleontology, geology, ecology, ontogeny, phylogeny, or systematics (collectively known as the "non-genomic record".

The methods of this invention extract information about function and function change by analyzing sequence data alone, and then by coupling this analysis with secondary, tertiary, and quaternary structural data. Those of ordinary skill in the art know, of course, of other sources of evoluionary information that does not come from genomic sequence data or crystal structures. These "non-genomic" data come from paleontology, geology, ecology, ontogeny, phylogeny, and systematics (collectively known as the "non-genomic record").

A. Correlating the topology of an evolutionary trees and the non-genomic record.

Conversely, and quite usefully, when a node in an evolutionary tree

Dates can be obtained approximately by protein sequence analysis. In cases where silent substitutions have not equilibrated, NED distances or other distances based on the analysis of silent codon substitutions can be used.

As discussed above, detailed analyses of evolutionary histories frequently can provide a solution to the most general problem of the conventional evolutionary paradigm, the difficulty in routinely identifying a homolog of a target sequence with known function within the database. By analysis of non-Markovian evolutionary behavior at the level of the protein, a model of secondary structure can be predicted. This prediction can be used in turn to detect long distance homologs in some cases and exclude the possibility of distant homology in others. This increases the likelihood that a homolog will be found with a known structure, behavior, or function for a new protein sequence. If one is found, then the logic associated with the conventional evolutionary paradigm can be applied to generate a hypothesis concerning the behavior or function of the protein.

The value of this post-genomic tool to assign behavior and structure to a target sequence problem is expected to grow over the near term, as the ratio of sequences supported by experimental studies to those not supported increases with the conclusion of genome projects, and as more sequences increase the detail of the evolutionary histories that can be extracted from the database directly, and therefore the quality of the predicted secondary structural model.

At the next level, analysis of non-Markovian behavior at the level of the gene can alert the biological chemist that the logic associated with the conventional evolutionary paradigm might not apply in individual cases. In particular, if an episode of rapid sequence evolution intervenes in the evolutionary tree between the sequence of interest and the sequence with the know behavior and function, the biological chemist is alerted to the possibility that the function of the protein might have changed. This alert is useful even with close homologs, as illustrated in the example with leptin.

But what if the evolutionary tree contains *no* protein with a sequence with assigned function, even one with low sequence similarity? Even with more limited evolutionary histories, post-genomic tools that analyze non-Markovian evolution at the level of the codon can be useful. By identifying the organisms that provide the sequences at the "leaves" of the evolutionary tree, it is frequently possible to correlate branches in the evolutionary tree with episodes in geological history, as determined from the fossil record. Especially in multicellular animals (metazoa), the fossil record can provide approximate dates for the emergence of new physiological function. In this case, it is possible to ask whether an episode of rapid sequence evolution in a protein family (in particular, an episode with a high expressed/silent ratio) occurred at the same time as a new physiological function emerged on earth. If so, a first level of hypothesis about physiological function can be proposed, even if no behavior or function of any kind is known for any of the modern proteins.

Perhaps the most transparent analysis of this type concerns proteins that underwent massive radiative divergences in metazoa approximately 600 million years ago. This is the time of the *Cambrian explosion*, an episode in terrestrial history that marks the massive radiative divergence of multicellular animals, including chordates. Proteins families undergoing rapid evolution at this time (for example, of protein tyrosine kinases and src homology 2 domains) are almost certainly involved in the basic processes by which multicellular animals develop from a single fertilized egg.

This type of analysis might be applied in the family of ribonuclease (RNase) A (E.C.2.7.7.16), a well known family of digestive proteins found in ruminants. The protein underwent rapid sequence evolution approximately 45 million years ago, a time where ruminant digestion emerged in mammals [T. M.

JERMANN, J. G. OPITZ, J. STACKHOUSE, J. and S. A. BENNER, Reconstructing the evolutionary history of the artiodactyl ribo nuclease superfamily. *Nature* **374**, 57-59 (1995).]. Thus, the rapid molecular evolution evident in the reconstructed evolutionary history of this protein suggests that the protein is important for ruminant digestive function.

B. Correlating features of patterns of evolution in specific branches in the evolutionary tree with the non-genomic record

This type of analysis is obviously strengthened if one adds now information concerning Ka/Ks values, covarion behavior, homoplasy, and compensatory changes.

C. Correlating evolutionary events in several protein families occuring at approximately the same time with the non-genomic record

This type of analysis can obviously contribute to the determination of pathways, interactions between proteins from different families. These hypotheses are ueful when designing two-hybrid systems, for example, to detect protein-protein contacts.

Use of non-stochastic behavior generally

One of ordinary skill in the art will recognize from Serial No. 07/857,224 that the methods of the instant invention view molecular evolution in a way quite distinct from the way in which standard tools analyze protein sequence data. Virtually all tools for comparing the sequences of homologous proteins assume a model for divergent evolution that is stochastic in outcome. This model treats a protein sequence as a linear string of letters, one letter for each amino acid. According to the model, each letter in the string changes (the gene and its corresponding protein mutates) at a rate that is independent of its position. According to the stochastic model, future and past mutations are independent. Mutations at one position are independent of mutations elsewhere.

Such a model is at best an approximation for the reality of protein evolution. In reality, proteins are not linear strings of letters. Rather, they are organic molecules that fold in three dimensions. In the folded form, some positions in a protein sequence are more easily mutatable (without destroying function) than others. Amino acids distant in the sequence but close in the fold frequently undergo correlated mutation. Future mutations are frequently not independent of past mutations. Thus, real proteins divergently evolving under functional constraints behave differently than expected based on the stochastic model.

The difference between the reality of divergent evolution of proteins that fold and expectation based on the stochastic model proves to be important, as was disclosed first in Serial No. 07/857,224. By comparing the patterns of substitution within a set of folded proteins undergoing divergent evolution with expectations for those patterns based on the stochastic model, one can extract information about the fold.

This makes the nuclear family more than a database organizational feature. Because the nuclear family holds a history of the pattern of divergent evolution under functional constraints in the protein, it holds information about the fold of the protein. From the sequences of proteins in the nuclear family alone, one can decide which amino acids lie on the surface of the folded structure, which lie inside, and which lie near the active site. Elements of secondary structure, the helices, strands, and loops can be identified. A model of tertiary structure can be built as well, all from the evolutionary history embodied in the nuclear family.

EXAMPLES

Example 1. Functional analysis of aromatase

Aromatase is a cytochrome P450-dependent enzyme that catalyzes a three step reaction that creates an estrogen from an androgen. The physiological consequences of estrogen biosynthesis in human biology are well known, even among laymen. Estrogen is also synthesized in primitive chordates such as *Amphioxus* (Callard et al., 1984), but not in other metazoans. Therefore, estrogen appears to have been invented as a hormone early in the divergent evolution of chordates, presumably by recruitment of steroids involved in developmental biology in more primitive metazoan ancestors.

Aromatase belongs to the cytochrome P450 superfamily of enzymes, which has some two dozen family members (Nebert et al., 1991). Members of the superfamily use a common chemical mechanism (Akhtar et al, 1997) to assimilate carbon, detoxify organic substances, and synthesize regulatory molecules. In biomedicine, variants of P450 oxidases can determine whether individuals have side effects to a therapeutic agent (Gonzalez & Nebert, 1990), and aromatase itself plays a significant role in the progression of some cancers.

Recent research has found remarkable complexity in the molecular biology of the aromatase gene family. Two aromatase genes are known in goldfish (Callard and Tchoudakova, 1997). In contrast, only a single gene is known in the horse (Boerboom et al., 1997), the rat (Hickey et al., 1990), the mouse (Terashima et al., 1991), the human (Harada, 1988), and the rabbit (Delarue et al, 1996). Both a functional gene and a pseudogene are found in oxen. The pseudogene is built from homologs of exons 2, 3, 5, 8, and 9 interspersed with a bovine repeat element (Fürbaß & Vanselow, 1995); it is transcribed but not translated. In several mammalian species, a single gene yields multiple forms of the mRNA for aromatase in different tissues via alternative splicing mechanisms. This is the case in humans (Simpson et al., 1997) and rabbits (Delarue et al. 1998).

A still different phenomenology is observed in the pig (*Sus scrofa*). Preliminary studies found three distinct mRNA molecules in different tissues with differences in their coding regions (Conley et al. 1996; Conley et al. 1997; Choi et al., 1996; Choi et al., 1997a; Choi et al., 1997b). It was suggested that these might have arisen from a single gene, possibly via RNA editing or alternative splicing (Conley et al. 1997).

Analogous collections of phenomenology are found throughout contemporary molecular biology for many molecular systems. "Why?" questions are often confounded by the complexity of the phenomenology. When "just so" stories are proposed, they need not be compelling, especially when they are supported by no evidence past the phenomena themselves.

One approach to obtain additional evidence to address functional questions in systems requires placing the molecular biological phenomena within an evolutionary context. To do this for the aromatases family, we began with experiments to determine whether the three mRNA isoforms (and the corresponding proteins) in pig arose through alternative splicing, via mRNA editing, or from distinct genes. PCR primers were designed from sequences located within the previously characterized exon 4 of the porcine aromatase type III gene (Choi et al., 1996, 1997a), a region that the cDNA studies suggested might have internal sequence differences (Choi et al., 1997a; Conley et al., 1997) and used to amplify pig genomic DNA. Initially, eight clones of the PCR products were sequenced. Four of these had the sequence corresponding to aromatase isoform I (ovarian type) as identified from cDNA, while four others had the sequence corresponding to aromatase isoform III (embryo type) as identified from cDNA.

With evidence that at least two aromatase genes could be found in pig genomic DNA, a restriction enzyme-based assay was designed to search genomic DNA in greater detail. *Nsi* I digests exon 4 from isoform I twice, and isoform III once. *Bsm* I digests exon 4 from isoform I once, but not exon 4 of isoform III. Exon 4 from isoform II (placental type) had no restriction sites for either enzyme. Restriction analysis of a total of 23 clones obtained from genomic DNA identified 8, 5, and 10 representatives of isoforms I, II, and III, respectively. No restriction digestion patterns indicative of a novel sequence were observed. Representative clones for isoforms I, II, and III were then sequenced. To further confirm the presence of exactly three aromatase isoforms within the porcine genome, primer pairs were designed from within the 5' and 3' junctions of exon 7. Sequence analysis of 10 clones derived from the PCR products identified six and four clones of isoforms II and III, respectively

With compelling evidence that the three variants of mRNA identified in cDNA studies arose from three paralogous genes (as opposed to editing or alternative splicing), we sought to place the paralogous genes within their historical context. Following standard tools to analyze protein sequences, pairwise alignments were constructed for the 136 pairs of proteins. An evolutionary distance (in PAM units) was calculated (with a variance) for each pair (Table 1). From this, an evolutionary tree was built for the mammalian sequences (Drawing 4), with branch lengths along internal nodes calculated to minimize a least squares distance were then constructed within the Darwin programming environment. The tree was adjusted to make the human and equine branchings consistent with paleontological records to obtain a "best consensus" tree. The sequences of the ancestral genes and proteins at branch points in the tree were then reconstructed. From there, mutations (including fractional mutations) at both the DNA level and protein level were assigned to individual branches in the tree using the method of Fitch (1971).

Based on the tree and the reconstructed evolutionary intermediates, K_a/K_s values were assigned to individual branches using the method of Li et al. (1985). These reflect the normalized ratio of substitutions at the level of the gene that change the encoded polypeptide sequence (non-synonymous substitutions) to substitutions at the level of the gene that do not change the encoded polypeptide sequence (synonymous substitutions). Lower K_a/K_s values generally reflect conservative episodes of evolution where function remains constant, while higher values frequently characterize episodes of evolution where function is changing (Trabesinger-Ruef et al., 1996; Messier & Stewart, 1997).

The average branch in the aromatase evolutionary tree has a value of K_a/K_s of 0.348. Inspection of the tree shows that the highest K_a/K_s values anywhere in the mammalian aromatase family (0.85 and 0.66) are found within the divergent evolution of the pig aromatases. These suggest that adaptive changes occurred during the triplication of the aromatase gene in pigs. Adaptive changes are well known to confuse simple models of molecular history built from standard sequence alignment and tree construction tools. Adaptive substitutions do not conform to stochastic rules modelling divergent evolution (Benner et al, 1997), do not accumulate in a clock-like fashion, and may arise through convergent and parallel evolution (Stewart et al., 1987).

Therefore, the evolutionary history of the aromatase family was re-analyzed using pairwise Neutral Evolutionary Distances (NEDs) (Liberles et al, 1999), obtained for the 136 pairs of aligned aromatase genes (Table 2). To estimate NEDs between the aromatase gene pairs, the number (*n*) of "2-fold redundant amino acids" (Cys, Asp, Glu, Phe, His, Lys, Asn, Gln, and Tyr) that are conserved in the aligned pairs was determined. The number of those amino acids that are encoded by the same codon (*c*) was then determined, and the fraction ($f2 = c/n$) of the codons that are the same is then tabulated (Table 2).

A variety of empirical studies show that the fixation of silent substitutions in conserved 2-fold redundant codon systems follows rate law that is a simple exponential "approach to equilibrium" $f2 = [0.5 \bullet \exp(-kt)] + 0.5$, where k is a single pseudo first order rate constant for transitions, and t is the time (Jukes & Cantor, 1969). The NED distance is defined by $NED_{x,y} = kt_{x,y} = \ln[(f2_{x,y}+0.5)/0.5]$.

The NED is a measure of evolutionary distance, not of evolutionary time. As distances, NEDs are additive, should obey the triangle inequality, and display other features that permit them to be used to build evolutionary trees, provided that k is constant over the period of evolutionary history being examined. A variety of empirical studies shows this to be approximately the case for many protein families. The approximation appears to be quite good for aromatase as well. Thus, if a fixed single lineage first order rate constant of 3×10^{-9} changes per base per year is assumed, the NED values indicate that fish and land vertebrates diverged 340 million years ago (mya), birds and mammals diverged 250 mya, primates and ungulates diverged 73 mya, horse and artiodactyls diverged 71 mya, and pigs and ruminants diverged 62 mya. Each of these dates is close to the date suggested by the paleontological record (Carroll, 1988).

The NED-based dating was used to assess two alternative models to explain the triplication of aromatase gene family in pigs. The first, advanced by Callard and Tchoudakova (1997), holds that the physiological specialization of aromatases through the formation of paralogs occurred early in vertebrate divergence, perhaps 400 mya, before fish and mammals diverged. If this were the case, then a functional explanation for the aromatase genes must be sought in fundamental features of vertebrate developmental biology, those that emerged early in vertebrate evolution. Conversely, the triplication of aromatase may occur in response to the domestication of pigs. In this case, a functional explanation for the aromatase genes would be found in the selective pressures applied by breeding programs.

The NEDs separating the three pig isoforms range from 0.154 (corresponding to a distance of 51 million years between the proteins) to 0.199 (corresponding to a distance of 66 million years). Recognizing that the total distances between two proteins are twice the distance along a single lineage from the point of divergence to the modern protein (half of the distance occurrs along one lineage after divergence, and half of the distance occurs along the other lineage), the NEDs suggest that the first duplication led to the three porcine aromatase genes occurred ca. 33 mya, and the second occurred ca. 25 mya. An evolutionary tree constructed from these NEDs is consistent with these conclusions, showing that the porcine aromatases branched after the lineage leading to pig diverged from the lineage leading to ox (Drawing 5). This tree shows a different branching order for the three porcine paralogs than the tree based on amino acid sequences, something not uncommon in the presence of substantial adaptive evolution. Nevertheless, the data are consistent with an evolutionary model that holds that the ancestor of pig and oxen (approximated in the fossil record most closely by the now extinct *Diacodexis* which lived perhaps 55 mya) contained a single aromatase gene, and that the paralogous genes in pig arose ca. 25 million years later. Thus, the paralogs in pig can be explained neither in terms of the fundamentals of vertebrate development, nor as a consequence of swine domestication.

Error in these dates can arise from two sources, standard error (which arises from fluctuation) and systematic error (which arises from the fact that the evolutionary model does not represent actual evolution). The first can be calculated by standard statistical approaches using standard statistical assumptions. The second cannot be calculated, as too little is known about possible systematic errors in the evolutionary model. The f2 distances are each based on ca. 120 two-fold redundant codon systems, and variances for the NEDs are given in Table 2. Inspection of the tree in Drawing 5 gives an indication of the actual error, as the NED between any ancestral sequences and all modern sequences derived from it should be the same. The calculated distance from the divergence of the three porcine enzymes to the type II enzyme is 31 million years, to isoform I is 32 million years, and to isoform III is 30 million years. Thus, the average reported (31 mya) could be as low as 30 and as high as 32 mya. All of these dates are in the Oligocene, after the first episode of cooling. The divergence of isoform I and III ranges from 24-26 mya. These apparent errors are less than the errors associated with the dating (from the fossil record) used to set the molecular clock.

Instead, an understanding of why pigs have three genes for aromatase must lie in the environment of (and events that occurred during) a time on Earth 25-33 mya. For this we turn to the paleontological, paleogeographical, and paleoclimatological records of that period, which is near the boundary between the Oligocene (38-25 mya) and the Miocene (25-5 mya), two epochs in the Cenozoic "Age of Mammals" (Prothero, 1994). This period is an unusual one in the history of the Earth. When characterized globally, the Earth during the Eocene (54 - 38 mya) was warm and tropical, evidently free of ice over the entire planet. By the end of the Eocene, however, the Earth had begun to suffer a dramatic cooling that was to lower the mean annual temperature by as much as 15 °C (Wolfe, 1978). Areas of the planet became covered with ice. And the impact of the cooling on the biosphere was dramatic. For example, perhaps 80% of the North American faunal genera became extinct (Prothero pp 113-114; Stucky, 1990). By the end of the Oligocene and into the Miocene 25 mya, however, the global cooling abated, the climate turned warmer, and the biosphere became more tropical.

Did this climate change occur in the environment where the ancestors of modern pigs were living just before the Oligocene-Miocene boundary? At this time, the North American and Eurasian fauna were geographically isolated. Modern peccaries (*Tayassuidae*), not pigs, emerged in the New World from ancestral suids that immigrated from Asia. North America cannot be the site for the triplication of the aromatase genes in pig, therefore, and its climate 25-33 mya is irrelevant to an explanation for the triplication of the aromatase genes in pigs.

Instead, modern pigs most likely emerged in Europe near the end of the Oligocene (Cooke & Wilkinson, 1978, but see also Pilgrim, 1941) from more primitive entelodonts such as *Archaeotherium*. During the Oligocene, the Dichobunids (the most probable ancestral stock) were most abundant in Europe. Likewise, the first true pig, *Propalaeochoerus*, from the late Oligocene, was common only in Europe (Cooke and Wilkinson, 1978; Carroll, 1988). This makes the paleoenvironment of Europe near the Oligocene-Miocene boundary relevant to the functional implications of the aromatase gene triplication in pigs.

Various paleobiological evidence suggests that the climate in Europe also deteriorated in the Oligocene and warmed in the Miocene. A study of amphibian distribution in the Oligocene of Europe, for example, is consistent with a significant drop of mean annual temperatures in the European Oligocene. In the Miocene, amphibians populations rebounded, corresponding to an improvement in the climate (Rocek, 1996). Likewise, analysis of the deer population suggested a subtropical climate returning to Europe in the early Miocene (Anzanza, 1993). The Iberian peninsula in the early Miocene had an intertropical to subtropical climate (Murelaga et al., 1999). Crocodiles also returned to Europe at the Oligocene-Miocene boundary (Antunes & Cahuzac, 1999). The presence of arboreal primates in the European Miocene also suggests a forested environment (Qi & Beard. 1998). Each of these facts (and many others) suggests that the second duplication of the aromatase gene in pigs occurred at the same time as the return of subtropical and warm temperate forests and woodlands to Europe, the type of environment for which suids are best adapted (Fortelius et al., 1996).

Immediately thereafter, the suids underwent a significant radiative divergence, and came to occupy all of the Old World. By the early Miocene, the two basal members that were to lead to all modern pigs, *Hyotherium* and *Xenochoerus*, were widespread in Europe, Asia, and Africa. The amelioration of the climate evidently assisted in this spread. For example, the pigs now in Africa apparently came from southwest Asia in the Early Miocene. A fossil of this date of a tetraconodontine pig has been reported from the Levant (van der Made & Tuna, 1999), through which the pigs would have migrated to get from Eurasia to Africa, and which was a tropical environment at the beginning of the Miocene (Tchernov, 1992). In the middle and late Miocene, modern suids had diversified in Europe in further response to the change in the paleoclimate (Fortelius et al., 1996).

Why might a change in climate with a return of forested (and perhaps tropical) ecosystems have led to a selection of pigs that had three different aromatase genes? We turned to porcine reproductive physiology for insight. We recently found that the type III aromatase was expressed by the embryo between day 11 and day 13 following fertilization, during the late pre-implantation period (Choi et al., 1997a,b). The estrogen generated by the type III isoform causes uterine undulation. This undulation, in turn, is expected to cause the spacing of the ca. 30 eggs that are fertilized in a typical conception, which eventually yield the 8-12 piglets that are normally birthed. In pigs, if the litter does not contain at least 5 individuals, the entire conception is aborted. Thus, the embryonic form of aromatase may have a role in spacing the embryos uniformly around the uterus, and preventing abortion. These are useful adaptations if one wants to have an increased litter size.

Evidence in the paleontological record suggests that the size of the litter in pigs increased dramatically 25-30 mya, at the same time as isoform III of aromatase was generated by triplication, the local paleoclimate warmed, and the pigs began a major radiative divergence. The ancestral suid *Archaeotherium*, disappearing from the fossil record at the end of the Oligocene, may have given birth to a single pup. All of the contemporary forms of pigs arising from the divergence of Hyotherium and Xenochoerus, known from the Early Miocene, have large litter sizes. Further, *Archaeomeryx*, the early Eocene artiodactyl that is presumed to be the ancestral ruminant, resembles the contemporary chevrotain, which also births a single pup.

The biogeography of the suids was again consulted to test the hypothesis that litter size increased in the suids near the time that the climate changed and the aromatase gene triplicated. As noted above, peccaries were isolated in the New World in the Early Oligocene, before the NED-derived date for the triplication of the aromatase gene in the Old World pigs. Consistent with the model, the peccary has only one offspring. The model predicts as well that the peccary should have only a single aromatase gene.

```
Pig
Type  I   C AAT CAT TAC ACG TGC CGA TTT GGC AGC AAA CTT GGG TTG GAA
              N   H   Y   T   C   R   F   G   S   K   L   G   L   E
    III   T AGT CAC TAC ACA TCC CGA TTT GGC AGC AAA CCT GGG TTG CAG
              S   H   Y   T   S   R   F   G   S   K   P   G   L   Q
     II   C AGT CAC TAC ACA TCC CGA TTC GGC AGC AAA CCT GGG TTG GAG
```

```
               S   H   Y   T   S   R   F   G   S   K   P   G   L   E
Peccary      C AGT CAC TAC ACA TCC CGA TTC GGC AGC AAA CCT GGG TTG CAG
               S   H   Y   T   S   R   F   G   S   K   P   G   L   Q

Pig
Type   I     TGC ATT GGC ATG CAT GAA AAA GGC ATC ATG TTT AAC AAT AA
              C   I   G   M   H   E   K   G   I   M   F   N   N   N
     III     TTC ATT GGC ATG CAT GAG AAA GGC ATT ATA TTC AAC AAT AA
              F   I   G   M   H   E   K   G   I   I   F   N   N   N
      II     TGC ATC GGC ATG TAT GAG AAG GGC ATC ATA TTT AAT AAT GA
              C   I   G   M   Y   E   K   G   I   I   F   N   N   D
Peccary      TTC ATT GGA ATG CAT GAG AAA GGC ATC ATA TTT AAC AAC AA
              F   I   G   M   H   E   K   G   I   I   F   N   N   N
```

To test this prediction, peccary seminal plasma (from the Center for Reproduction of Endangered Species, Zoological Society of San Diego) was subjected to PCR amplification using exon 4-specific primers as described above. Bands having the expected sizes were observed by agarose gel electrophoresis. Five clones derived from the PCR products were found to have identical sequences, all different from the sequences of the pig aromatase. The NED comparison (using a rate constant of 3 x 10^{-9} changes per base per year) suggested that the peccary diverged 40 mya from the pig, corresponding to the fossil record and the known isolation of the New and Old World paleoecosystems.

The molecular biological, fossil, paleoecological, and physiological evidence are all consistent with a model that proposes that climate changes in Europe at the end of the Oligocene selected for pigs that had larger litter sizes. The successful lineage generated a new embryo aromatase by gene duplication, and expressed it at the time of implantation, forming the molecular basis of the physiology that enabled large litter sizes. It is possible to speculate on why a conversion from an open, savannah like environment to a forested environment might enable larger litter sizes. Contemporary savannah babies are large and born with the ability to run, presumably because hiding is no alternative. In contrast, in a forested environment, pups are easier to hide, permitting them to be smaller and less precocious at birth, permitting in turn a larger number of pups for the same total birth weight. Indeed, the contemporary *Sus scrofa* sow hides her piglets in earthen hollows covered with leaves (Eisenberg, 1981).

Implantation is one of the least well understood steps in mammalian reproductive biology, including human reproductive biology. Implantation is, of course, found only in mammal reproductive physiology, and is itself therefore a relatively recent innovation in physiology, emerging perhaps 200 million years ago. This analysis emphasizes the degree of innovation and experimentation that is continuing in mammalian reproductive physiology. Further, the analysis is a combination of computational informatics, geology, paleontology, physiology, molecular biology and chemistry. Analogous analyses should be applicable in functional genomics throughout the biological, biomedical and biochemical sciences, especially as genome projects are completed and as new tools become available to analyze genomic databases.

References for Example 1

Akhtar, M., LeeRobichaud, P., Akhtar, M.E., Wright, J.N. (1997) The impact of aromatase mechanism on other P450s. *J. Steroid Biochem. Mol. Biol.* **61**, 127-132.

Antunes, M. T., Cahuzac, B. (1999) Crocodilian faunal renewal in the Upper Oligocene of Western Europe. *Comptes Rend. L'Acad. Sci. Serie II Fascicule A-Sci. Terre Planetes.* **328**, 67-72.

Azanza, B. (1993) Systematics and evolution of the genus *Procervulus* (Cervidae, Artiodactyla, Mammalia) of the lower Miocene of Europe. *Comptes Rend. L'Acad. Sci. Serie II.* **316**, 717-723.

Benner, S. A., Cannarozzi, G., Chelvanayagam, G., Turcotte, M. (1997) *Bona fide* predictions of protein secondary structure using transparent analyses of multiple sequence alignments. *Chem. Rev.* **97**, 2725-2843.

Benner, S. A., Trabesinger-Ruef, N., Schreiber, D. R. (1998) Exobiology and post-genomic science. Converting primary structure into physiological function. *Adv. Enzyme Regul.* **38**, 155-180.

Boerboom, D., Kerban, A., Sirois, J. (1997) Molecular characterization of the equine cytochrome P450 aromatase cDNA and its regulation in preovulatory follicles. *Biol. Reprod..* **56**, 479-479, Suppl. 1.

Buck, C. D. (1988) *A Dictionary of Selected Synonyms in the Principal European Languages.* Chicago, University of Chicago Press, Paperback ed., p. 160.

Callard, G. V., Tchoudakova, A. (1997) Evolutionary and functional significance of two CYP19 genes differentially expressed in brain and ovary of goldfish. *J. Steroid Biochem. Mol. Biol.* **61**, 387-392.

Callard, G. V., Pudney, J. A., Kendall, S. L., Reinboth, R. (1984) In vitro conversion of androgen to estrogen in *Amphioxus* gonadal tissues. *Gen. Comp. Endocrinol.* **56**, 53-58.

Carroll, R. L. (1988) *Vertebrate Paleontology and Evolution.* N.Y., Freeman.

Chang, X. T., Kobayashi, T., Kajiura, H., Nakamura, M., Nagahama, Y. (1997) Isolation and characterization of the cDNA encoding the tilapia (*Oreochromis niloticus*) cytochrome P450 aromatase (P450arom), Changes in P450arom mRNA, protein and enzyme activity in ovarian follicles during oogenesis. *J. Mol. Endocrinol.* **18**, 57-66.

Choi, I., Collante, W. R., Simmen, R. C. M., Simmen, F. A. (1997a) A developmental switch in expression from blastocyst to endometrial/placental-type cytochrome p450 aromatase genes in the pig and horse. *Biol. Reprod.* **56**, 688-696.

Choi, I.H., Troyer, D. L., Cornwell, D. L., Kirby-Dobbels, K. R., Collante, W. R., Simmen, F. A. (1997b) Closely related genes encode developmental and tissue isoforms of porcine cytochrome P450 aromatase. *DNA Cell. Biol.* **16**,769-777.

Choi, I., Simmen, R. C. M., Simmen, F. A. (1996) Molecular cloning of cytochrome P450 aromatase complementary deoxyribonucleic acid from periimplantation porcine and equine blastocysts identifies multiple novel 5'-untranslated exons expressed in embryos, endometrium, and placenta. *Endocrinol.* **137**, 1457-1467.

Colbert, E. H. (1941) The osteology and relationships of *Archaeomeryx*, an ancestral ruminant. *Amer. Mus. Novit.* **1135**, 1-24.

Conley, A., Corbin, J., Smith, T., Hinshelwood, M., Liu, Z., Simpson, E. (1997) Porcine aromatases, studies on tissue-specific functionally distinct isozymes from a single gene? *J. Steroid Biochem. Mol. Biol.* **61**, 407-413.

Conley, A. J., Corbin, C. J., Hinshelwood, M. M., Liu, Z., Simpson, E. R., Ford, J. J., Harada, N. (1996) Functional aromatase expression in porcine adrenal gland and testis. *Biol Reprod.* **54**,497-505.

Cooke, H. B. S., Wilkinson, A. F. (1978) *Suidae* and *Tayassuidae,* in *Evolution of African Mammals*, V. J. Maglio and H. B. S. Cooke, eds. Cambridge, Harvard University Press, 438-482.

Delarue, B., Breard, E., Mittre,H., Leymarie, P. (1998) Expression of two aromatase cDNAs in various rabbit tissues. *J. Steroid Biochem. Mol. Biol.* **64**, 113-119.

Delarue, B., Mittre, H., Feral, C., Benhaim, A., Leymarie, P. (1996) Rapid sequencing of rabbit aromatase cDNA using RACE PCR. *Comptes Rend. L'Acad. Sci. Serie III Sciences De La Vie-Life Sciences* **319**,663-670.

Eisenberg, J. F. (1981) *The Mammalian Radiations. An Analysis of Trends in Evolution, Adaptation, and Behavior.* Chicago, Univ. Chicago Press, p 196.

Fitch, W. (1971) Towards defining the course of evolution. Minimum change for a specific tree topology. *Syst. Zoology* **20**, 406-416.

Fortelius, M., van der Made, J., Bernor, R. L. (1996) Middle and Late Miocene Suoidea of Central Europe and the Eastern Mediterranea, Evolution, Biogeography and Paleoecology. in *The Evolution of Western Eurasian Neogene Mammal Fanas.* R. L. Bernor, V. Fahlbusch, and H.-W. Mittmann eds. Columbia Univ. Press, 348-377.

Fürbaβ R, Vanselow J. (1995) An aromatase pseudogene is transcribed in the bovine placenta. *Gene* **154**,287-291.

Gonnet, G. H., Benner, S. A. (1991) Computational Biochemistry Research at ETH. *Technical Report 154, Departement Informatik,* March, 1991.

Gonzalez, F. J., Nebert, D. W. (1990) Evolution of the P450-gene superfamily. Animal plant warfare, molecular drive and human genetic-differences in drug oxidation. *Trends Genet.* **6**, 182-186.

Harada, N. (1988) Cloning of a complete cDNA encoding human aromatase, immunochemical identification and sequence analysis. *Biochem. Biophys. Res. Comm.* **156**, 725-732.

Hickey, G. J., Krasnow, J. S., Beattie, W. G., Richards, J. S. (1990) Aromatase cytochrome P450 in rat ovarian granulosa cells before and after luteinization. Adenosine 3',5'-monophosphate-dependent and independent regulation. Cloning and sequencing of rat aromatase cDNA and 5' genomic DNA. *Mol. Endocrinol.* **4**, 3-12.

Hinshelwood, M. M., Corbin, C. J., Tsang, P. C. and Simpson,E.R. (1993) Isolation and characterization of a complementary deoxyribonucleic acid insert encoding bovine aromatase cytochrome P450. *Endocrinology* **133**, 1971-1977.

Jukes, T. H., Cantor, C. R. (1969) Evolution of proteins molecules. in *Mammalian Protein Metabolism*, H. N. Munro, ed. N.Y. Academic Press, pp. 21-123.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111-120.

Li, W.-H., Wu, C.-I., Luo, C.-C. (1985) A new method for estimating synonymous and non-synonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2**, 150-174.

Liberles, D. A., Caraco, M. D., Benner, S. A. (1999) Using neutral evolutionary distances to estimate the dates of divergence of proteins. *in preparation.*

McPhaul, M.J., Noble, J.F., Simpson, E.R., Mendelson, C.R., Wilson, J.D. (1988) The expression of a functional cDNA encoding the chicken cytochrome P-450-arom (aromatase) that catalyzes the formation of estrogen from androgen. *J. Biol. Chem.* **263**, 16358-16363.

Messier, W., Stewart, C. B. (1997) Episodic adaptive evolution of primate lysozymes (1997) *Nature* **385**,151-154.

Murelaga, X., de Broin, F. D., Suberbiola, X. P., Astibia, H. (1999) Two new chelonian species from the Lower Miocene of the Ebro Basin (Bardenas Reales of Navarre). *Comptes Rend. L'Acad. Sci. Serie II Fascicule A-Sci. Terre Planetes.* **328**, 423-429.

Nebert, D. W., Nelson, D. R., Coon, M. J., Estabrook, R. W., Feyereisen, R., Fujiikuriyama, Y., Gonzalez, F. J., Guengerich, F. P., Gunsalus, I. C., Johnson, E.F., Loper, J. C., Sato, R., Waterman, M. R., Waxman, D. J. (1991) The P450 superfamily. Update on new sequences, gene-mapping, and recommended nomenclature. *DNA Cell Biol.* **10**,1-14.

Pilgrim, G. E. (1941) The dispersal of the Artiodactyla, *Biol. Rev.*, **16**, 134-163.

Prothero, D. R. (1994) *The Eocene-Oligocene Transition, Paradise Lost* NY, Columbia Univ. Press.

Qi, T., Beard, K. C. (1998) Late Eocene sivaladapid primate from Guangxi Zhuang Autonomous Region, People's Republic of China. *J. Human Evol.* **35**, 211-220.

Rocek, Z. (1996) The salamander *Brachycormus noachicus* from the Oligocene of Europe, and the role of neoteny in the evolution of salamanders. Palaeontology 39, 477-495.

Rose, K. D. (1982) Skeleton of *Diacodexis*, oldest known artiodactyl. *Science* **236**, 621-623.

Savage, R. J. G., Long M. R. (1986) Mammal Evolution. An Illustrated Guide. N.Y., Facts on File Publ., p 213.

Scott, W. B. (1937) *A History of Land Mammals in the Western Hemisphere*. N.Y. McMillan.

Shen, P., Campagnoni, C.W., Kampf, K., Schlinger, B.A., Arnold, A.P., Campagnoni, A.T. (1994) Isolation and characterization of a zebra finch aromatase cDNA. *In situ* hybridization reveals high aromatase expression in brain. *Brain Res. Mol. Brain Res.* **24**, 227-237.

Simpson, E. R., Michael, M. D., Agarwal, V. R., Hinshelwood, M. M., Bulun, S. E., Zhao, Y. (1997) Expression of the CYP19 (aromatase) gene. An unusual case of alternative promoter usage. *FASEB J.*, **11**, 29-36.

Stewart, C. B., Schilling, J. W., Wilson, A. C. (1987) Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature* **330**, 401-404.

Stucky, R. K. (1990) Evolution of land mammal diversity in North America during the Cenozoic. *Curr. Mammalogy* **2**, 375-432.

Tanaka, M., Fukada, S., Matsuyama, M., Nagahama,Y. (1995) Structure and promoter analysis of the cytochrome P-450 aromatase gene of the teleost fish, medaka (*Oryzias latipes*). *J. Biochem.* **117**, 719-725.

Tchernov, E. (1992) The Afro-Arabian component in the levantine mammalian fauna. A short biogeographical review. *Israel J. Zoology* **38**, (3-4) 155-192.

Terashima, M., Toda, K., Kawamoto, T., Kuribayashi, I., Ogawa, Y., Maeda, T., Shizuta,Y. (1991) Isolation of a full-length cDNA encoding mouse aromatase P450. *Arch. Biochem. Biophys.* **285**, 231-237.

Trabesinger-Ruef, N., Jermann, T. M., Zankel, T. R., Durrant, B., Frank, G., Benner, S. A. (1996) Pseudogenes in ribonuclease evolution. A source of new biomacromolecular function? *FEBS Lett.* **382**, 319-322.

Trant, J. M. (1994) Isolation and characterization of the cDNA encoding the channel catfish (*Ictalurus punctatus*) form of cytochrome P450arom. *Gen. Comp. Endocrinol.* **95**, 155-168.

van der Made J, Tuna V. (1999) A tetraconodontine pig from the Upper Miocene of Turkey. *Trans. Royal Soc. Edinburgh. Earth Sci.* **89**, 227-230.

Wolfe, J. A. (1978) A paleobotanical interpretation of Tertiary climates in the Northern Hemisphere. *American Sci.* **66**, 694-703.

Example 2. Covarion behavior

Functional changes leave signatures in the patterns of sequence evolution in a protein family. Covarion behavior was detected in alcohol dehydrogenase [Ben89] and superoxide dismutase [Miy95]. As a preliminary study in the past year, we examined elongation factors (EF). These are proteins that have diverged far more slowly; indeed, they are archetypal examples of a protein that performs the "same" function in all three kingdoms of life.

In the study, thirty EF-Tu/EF-1α protein sequences were aligned over 380 sites using the alignment program DARWIN. Replacement rates per site for bacterial and eukaryotic EFs were estimated using a gamma-based, maximum likelihood (ML) model for protein sequences (JTT + Γ) and the phylogeny of Baldauf *et al.* [Bal96] for EF-Tu and EF-1α. An α of 0.78 was calculated for the entire tree, with a standard deviation (SD) of 0.05 using parametric bootstrapping (evolutionary simulations) [Swo96]. Interestingly, the α values for the bacterial and eukaryotic subtrees were significantly different from that for the entire tree [0.46 (0.04) and 0.38 (0.04), respectively]. These reductions in α for bacteria and eukaryotes alone are expected of a non-stationary covarion process.

The distribution of rate differences per site between bacterial and eukaryotic EFs is leptokurtotic; i.e., over- and under-represented in the mean and tails versus "shoulders," respectively, relative to the expectations of a normal distribution. Thirty seven percent of the sites have essentially the same rate in the two groups (rate difference of ~0), as expected under a stationary gamma process. However, 18 and 21 sites evolve >2 SD faster in bacteria than eukaryotes, and vice versa, respectively. These 10% of the sites are most responsible for the covarion characteristics of EF-Tu and EF-1α.

Residues displaying abnormal evolutionary behavior were then mapped to a three dimensional model of the protein based on a crystal structure of ET-Tu. These were used to generate structural hypotheses for the different behavioral differences that were known. For example, bacterial EF-Tu binds GDP ~100 fold tighter than GTP. Eukaryotic EF-1α, in contrast, binds both with similar affinities. EF-Tu regenerates its active form by binding to the single-subunit nucleotide exchange factor EF-Ts. EF-1α requires the multi-subunit nucleotide exchange factor EF-1βγδ. EF-1α also interacts with the cytoskeleton and may thereby play a role in cellular transformation and apoptosis. EF-Tu can have no such role in bacteria. Residues were identified that, at the level of hypothesis, are responsible for each of these behavioral differences.

Covarion behavior indicates changing function. It is therefore expected to correlate positively with events with high K_a/K_s ratios. We will see if that is correct. Because K_a/K_s ratios use a silent substitution clock that ticks rapidly, while covarion analysis does not, the two are somewhat complementary. This addresses another of the concerns of the referees, who objected that K_a/K_s ratios were not applicable far enough back in time for their tastes (only ca. 500 my).

Example 3. Identifying mutations and in vitro properties of seminal ribonuclease that contribute to selected function.

Bovine seminal ribonuclease (RNase) diverged from bovine pancreatic RNase approximately 35 million years ago. Seminal RNase represents approximately 2% of the total protein in bovine seminal plasma. It displays antispermatogenic activity [Dostal, J., Matousek, J. (1973) Isolation and some chemical properties of aspermatogenic substance from bull seminal vesicle fluid. *J. Reprod. Fertil.* **33**, 263-274],immunosuppressive activity [Soucek, J., Matousek, J. (1981) Inhibitory effect of bovine seminal ribonuclease on activated lymphocytes and lymphoblastoid cell lines in vitro. *Folia Biol. Praha* **27**, 334-345. Soucek, J., Hrubá, A., Paluska, E., Chudomel, V., Dostál, J., Matousek, J. (1983) Immunosuppressive effects of bovine seminal fluid fractions with ribonuclease activity. *Folia biologica (Praha)* **29**, 250-261. Soucek, J., Chudomel, V., Potmesilova, I., Novak, J. T. (1986) Effect of ribonucleases on cell, mediated lympholysis reaction and on GM, CFC colonies in bone marrow culture. *Nat. Immun. Cell Growth Regul.* **5**, 250-258], and cytostatic activity against many transformed cell lines [Matousek, J. (1973) The effect of bovine seminal ribonuclease on cells of Crocker tumor in mice. *Experientia* **29**, 858. Vescia, S., Tramontano, D., Augusti-Tocco, G., D'Alessio, G. (1980) In vitro studies on selective inhibition of tumor cell growth by seminal ribonuclease. *Cancer Res.* **40**, 3740] Each of these biological activities is essentially absent from pancreatic RNase. Further, seminal RNase binds to anionic glycolipids, binds and melts duplex DNA, hydrolyzes duplex RNA, has a dimeric quaternary structure, and binds to spermatozoa.

Each of these behaviors is measured *in vitro* and is well known in the art. In the absence of the method of the instant invention, the behaviors are difficult to interpret. Some, any, or all of the behaviors might serve an adaptive role. It is possible that none of these behaviors serve adaptive roles. Indeed, it is conceivable that the protein has no adaptive role at all. This makes it difficult to make even the simplest research decisions, as the only *in vitro* properties of a protein that are interesting to study are those that have a physiological function.

To resolve these issues, genes for seminal and pancreatic RNases were obtained from a variety of organisms closely related to *Bos taurus*, using cloning procedures well known in the art. These were then sequenced, and a maximum parsimony tree was constructed using MacClade. From this tree were calculated the sequences of RNases that were intermediates in the evolution of the seminal RNase, using the maximum parsimony method well known in the art.

Next, the ratio of expressed to silent substitutions was calculated along each branch of the evolutionary tree. A very high ratio of expressed to silent substitutions was observed in the evolutionary period following the divergence of kudu [Trabesinger-Rüf, N., Jermann, T. M., Zankel, T. R., Durrant, B., Frank, G., Benner. S. A. Pseudogenes in ribonuclease evolution. A source of new biomacromolecular function? *FEBS Lett.* **382**, 319-322 (1996)] from the lineage leading to ox, until the divergence of water buffalo and ox. This is indicative of an episode of adaptive evolution, where the protein acquires a new physiological function. Further work indicated that the seminal RNase gene was not expressed in the period of evolution since the divergence of the seminal RNase family and the divergence of kudu.

Last, protein engineering methods were used to prepare the seminal RNase that was at the beginning of the episode of rapid sequence evolution. It properties were then examined experimentally. It was discovered that the ability of the protein to bind to anionic glycolipids was roughly the same before and after this episode of rapid evolution. So too was its sensitivity to inhibition by placental RNase inhibitor. Thus, both of these properties are not likely to be under selective pressure.

In contrast, the immunosuppressivity of the ancestral RNase (IC_{50} ca. 8 micrograms/mL) was greater than that of pancreatic RNase (IC_{50} ca. 100 micrograms/mL). But following the period of rapid sequence evolution characteristic of a protein evolving to serve a new physiological function, the immunosuppressivity became still greater (IC_{50} ca. 2 micrograms/mL). Thus, one concludes that immunosuppressivity as measured *in vitro* is a selected trait of the protein, or is closely structurally coupled to a trait that is selected.

Likewise, the ability of the seminal RNase protein to bind and melt duplex DNA, and to hydrolyze duplex RNA, also underwent rapid increase between the time of divergence of kudu from modern ox. Thus, it too is either a selected trait of the protein, or is closely structurally coupled to a trait that is selected.

In vitro experiments in biological chemistry extract data on proteins and nucleic acids (for example) that are removed from their native environment, often in pure or purified states. While isolation and purification of molecules and molecular aggregates from biological systems is an essential part of contemporary biological research, the fact that the data are obtained in a non-native environment raises questions concerning their physiological relevance. Properties of biological systems determined *in vitro* need not correspond to those *in vivo*, and properties determined *in vitro* need have no biological relevance *in vivo*.

To date, there has been no simple way to say whether or not biological behaviors are important physiologically to a host organism. Even in those cases where a relatively strong case can be made for physiological relevance (for example, for enzymes that catalyze steps in primary metabolism), it has proven to be difficult to decide whether individual properties of that enzymes (k_{cat}, K_m, kinetic order, stereospecificity, etc.) have physiological relevance. Especially difficult, however, is to ascertain which behaviors measures *in vitro* play roles in "higher" function in metazoa, including digestion, development, regulation, reproduction, and complex behavior.

Analysis of non-Markovian behavior, as described above, permits the biological chemist to identify episodes in the history of a protein family where new function is emerging. This suggests a general method to determine whether a behavior measured *in vitro* is important to the evolution of new physiological function. We may take the following steps:

(a) Prepare in the laboratory proteins that have the reconstructed sequences corresponding to the ancestral proteins before, during, and after the evolution of new biological function (34), as revealed by an episode of high expressed to silent ratio of substitution in a protein. This high ratio compels the conclusion that the protein itself serves a physiological role, one that is changing during the period of rapid non-Markovian sequence evolution.

(b) Measure in the laboratory the behavior in question in ancestral proteins before, during, and after the evolution of new biological function, as revealed by an episode of high expressed to silent ratio of substitution. Those behaviors that increase during this episode are deduced to be important for physiological function. Those that do not are not.

An example of this method was applied to the bovine seminal ribonuclease (RNase) family. Bovine seminal RNase diverged from bovine pancreatic RNase approximately 35 million years ago. Seminal RNase represents approximately 2% of the total protein in bovine seminal plasma. It displays antispermatogenic activity [J. Dostal and J. Matousek, Isolation and some chemical properties of aspermatogenic substance from bull seminal vesicle fluid. *J. Reprod. Fertil.* **33**, 263-274 (1973).], immunosuppressive activity [J. Soucek, Matousek, J., Inhibitory effect of bovine seminal ribonuclease on activated lymphocytes and lymphoblastoid cell lines in vitro. *Folia Biol. Praha* **27**, 334-345 (1981)., J. Soucek, A. Hrubá, E. Paluska, V. Chudomel, J. DostáL and J. Matousek, Immunosuppressive effects of bovine seminal fluid fractions with ribonuclease activity. *Folia biologica (Praha)* **29**, 250-261 (1983)., J. Soucek, V. Chudomel, I. Potmesilova, and J. T. Novak, Effect of ribonucleases on cell, mediated lympholysis reaction and on GM, CFC colonies in bone marrow culture. *Nat. Immun. Cell Growth Regul.* **5**, 250-258 (1986)], and cytostatic activity against many transformed cell lines [J. Matousek, The effect of bovine seminal ribonuclease on cells of Crocker tumor in mice. *Experientia* **29**, 858-859 (1973), S. Vescia, D. Tramontano, G. Augusti-Tocco and G. D'alessio, In vitro studies on selective inhibition of tumor cell growth by seminal ribonuclease. *Cancer Res.* **40**, 3740-3744 (1980)] Each of these biological activities is essentially absent from pancreatic RNase. Further, seminal RNase binds to anionic glycolipids, binds and melts duplex DNA, hydrolyzes duplex RNA, has a dimeric quaternary structure, and binds to spermatozoa.

Each of these behaviors is measured *in vitro*, as is the case for a wide range of biological phenomenology recorded in the literature. The behaviors are difficult to interpret. Some, any, or all of the behaviors might serve an adaptive role. It is possible that none of these behaviors serve adaptive roles. Indeed, it is conceivable that the protein has no adaptive role at all. This makes it difficult to make even the simplest research decisions, as the only *in vitro* properties of a protein that are interesting to study are those that have a physiological function.

To resolve these issues using the post-genomic method outlined above, genes for seminal and pancreatic RNases were obtained from a variety of organisms closely related to *Bos taurus*, using cloning procedures well known in the art. These were then sequenced, and a maximum parsimony tree was constructed using MacClade. From this tree were calculated the sequences of RNases that were intermediates in the evolution of the seminal RNase, using the maximum parsimony method and checked using maximum likelihood tools implemented in Darwin (23).

Next, the ratio of expressed to silent substitutions was calculated along each branch of the evolutionary tree. A very high ratio of expressed to silent substitutions was observed in the evolutionary period following the divergence of cape buffalo [N. Trabesinger-RüF, T. M. Jermann, T. R. Zankel, B. Durrant, G. Frank and S. A. Benner, Pseudogenes in ribonuclease evolution. A source of new biomacromolecular function? *FEBS Lett.* **382**, 319-322 (1996).] from the lineage leading to ox, until the divergence of water buffalo and ox. This is indicative of an episode of adaptive evolution, where the protein acquires a new physiological function. Further work indicated that the seminal RNase gene was not expressed in the period of evolution since the divergence of the seminal RNase family and the divergence of cape buffalo.

Last, protein engineering methods were used to prepare the seminal RNase that existed at the beginning of the episode of rapid sequence evolution. Its properties were then examined experimentally. It was discovered that the ability of the protein to bind to anionic glycolipids was roughly the same before and after this episode of rapid evolution. So too was its sensitivity to inhibition by placental RNase inhibitor. Thus, both of these properties are not likely to be under selective pressure.

In contrast, the immunosuppressivity of the ancestral RNase (IC_{50} ca. 8 micrograms/mL) was greater than that of pancreatic RNase (IC_{50} ca. 100 micrograms/mL) (J. Sleasman, M. Rojas, personal communication). But following the period of rapid sequence evolution characteristic of a protein evolving to serve a new physiological function, the immunosuppressivity became still greater (IC_{50} ca. 2 micrograms/mL). Thus, one concludes that immunosuppressivity as measured *in vitro* is a selected trait of the protein, or is closely structurally coupled to a trait that is selected.

Likewise, the ability of the seminal RNase protein to bind and melt duplex DNA, and to hydrolyze duplex RNA, also underwent rapid increases between the time of divergence of cape buffalo from modern ox. Thus, it too is either a selected trait of the protein, or is closely structurally coupled to a trait that is selected. In contrast, dimeric structure did not emerge during this period. Dimeric structure, therefore, is presumably not as important to the new selected function of the protein, although it may be a trait that was initially useful in the selection of the system for further optimization during the period of rapid evolution.

Example 4. Assignment of episodes of adaptive evolution in the protein leptin, and placing these in predicted secondary structural elements

From the GenBank database, DNA and protein sequences were retrieved for the genes encoding leptins and the corresponding proteins, also known as the obesity gene product. A multiple alignment for the protein sequences was constructed for the DNA sequences and the protein sequences. These were converted to a file suitable for MacClade to use. For both the DNA and protein sequences, a tree using MacClade was built based on the known relationship between the organisms from which these sequences were derived; this proved to be the most parsimonious tree as well. MacClade was also used to built a tree for the protein sequences based on the known relationship between organisms; this proved *not* to be the most parsimonious tree (by 1 change). The DNA tree was taken to be definitive because of its consistency with the biological (cladistic) data showing that the primates form a clade.

A secondary structure prediction was made for the protein family using the tools disclosed in Serial No. 07/857,224. The evolutionary divergence of the sequences available for the leptin family is small; only 21 PAM units (point accepted mutations per 100 amino acids), predictions were biased to favor surface assignments [Benner, S. A., Badcoe, I., Cohen, M. A., Gerloff, D. L. *Bona fide* prediction of aspects of protein conformation. Assigning interior and surface residues from patterns of variation and conservation in homologous protein sequences. *J. Mol. Biol.* **235,** 926-958 (1994)]. Thus, positions holding conserved KREND were assigned as surface residues, conserved H and Q were assigned to the surface as well, while positions holding conserved CST were assigned as uncertain. suface and interior assignments are summarized in Table 3.

A secondary structure was then predicted for the leptins using the methods disclosed in Serial No. 07/857,224. The multiple alignment is shown in Table 3. Five separate secondary structural elements were identified results are summarized in Table 3. A disulfide bond is presumed to connect positions 96 and 146. These secondary structural elements can be accommodated by only a small number of overall folds. Interestingly, the pattern of secondary structure in this prediction is consistent with an overall fold that resembles that seen in cytokines such as colony stimulating factor [Hill, C., P., Osslund, T. D., Eisenberg, D. (1993) *Proc. Nat. Acad. Sci.* **90**, 5176-5181] and human growth hormone [de Vos, A. M., Ultsch, M. & Kossiakoff, A. A. (1992). *Science* **255**, 306-312].

To decide whether evolutionary function may have changed under selective pressure during the divergent evolution of the protein family, a multiple alignment of the protein sequences and a multiple alignment for the corresponding DNA sequences were constructed. A MacClade-generated maximum parsimony tree was printed for each position in the protein sequence where there was a change, and for each position in the DNA sequence where there was a change. Each mutation on each tree was examined by hand, and silent and expressed mutations occurred were assigned to individual branches on the evolutionary tree. For each branch of the tree, the sum of the number of silent and expressed changes were tabulated, and the ratio of expressed to silent changes calculated. These are shown in Drawing 1. Tables 4 and 5 contain the data used in this example.

The branches on the evolutionary tree leading to the primate leptins from their ancestors at the time that rodents and primates diverged had an extremely high ratio of expressed to silent changes. From this analysis, it was concluded that the biological function of leptins has changed significantly in the primates rlative to the function of the leptin in the common ancestor of primates and rodents.

This approach can be illustrated in a biomedically interesting family of proteins by examining the protein leptin, a protein whose mutation in mice is evidently correlated with obesity, and was previously known as the "obesity gene protein". The protein has attracted substantial interest in the pharmaceutical industry, especially after a human gene encoding a leptin homolog was isolated. According to the conventional evolutionary paradigm, because it is a homolog of the mouse leptin, the human leptin must also play a role in obesity, and might be an appropriate target for pharmaceutical companies seeking human pharmaceuticals to combat this common condition in the first world.

DNA and protein sequences were retrieved for the genes encoding leptins. A multiple alignment for the protein sequences was constructed for the DNA sequences and the protein sequences. Congruent tress for both the DNA and protein sequences were then constructed, and sequences at the nodes of the tree reconstructed using MacClade [W. P. Maddison, D. R. Maddison, *MacClade. Analysis of Phylogeny and Character Evolution*. Sinauer Associates, Sunderland MA (1992).] and the known relationship between the organisms from which these sequences were derived. For the DNA sequences, the biologically most plausible tree proved to be the most parsimonious tree as well. The most parsimonious tree for the protein sequences proved *not* to be the most plausible tree (by one change) from a biological perspective. The DNA tree was taken to be definitive because of its consistency with the biological (cladistic) data.

A secondary structure prediction was made for the protein family. The evolutionary divergence of the sequences available for the leptin family is small - only 21 PAM units (point accepted mutations per 100 amino acids) - and predictions were biased to favor surface assignments [S. A. Benner, I. Badcoe, M. A. Cohen and D. L. Gerloff, *Bona fide* prediction of aspects of protein conformation. Assigning interior and surface residues from patterns of variation and conservation in homologous protein sequences. *J. Mol. Biol.* **235,** 926-958 (1994).]. Thus, positions holding conserved KREND were assigned as surface residues, conserved H and Q were assigned to the surface as well, while positions holding conserved CST were assigned as uncertain.

Five separate secondary structural elements were identified. A disulfide bond was presumed to connect positions 96 and 146. These secondary structural elements can be accommodated by only a small number of overall folds. Interestingly, the pattern of secondary structure in this prediction is consistent with an overall fold that resembles that seen in cytokines such as colony stimulating factor [C. P. Hill, T. D. Osslund and D. Eisenberg, The structure of granulocyte colony stimulating factor and its relationship to other growth factors. *Proc. Nat. Acad. Sci.* **90**, 5176-5181 (1993).] and human growth hormone [A. M. De Vos, M. Ultsch and A. A. Kossiakoff, Human growth-hormone and extracellular domain of its receptor. Crystal-structure of the complex.*Science* **255**, 306-312 (1992).].

To decide whether evolutionary function may have changed under selective pressure during the divergent evolution of the protein family, silent and expressed mutations were assigned to individual branches on the evolutionary tree. For each branch of the tree, the sum of the number of silent and expressed changes were tabulated, and the ratio of expressed to silent changes calculated. These are shown in Drawing 2.

The branches on the evolutionary tree leading to the primate leptins from their ancestors at the time that rodents and primates diverged had an extremely high ratio of expressed to silent changes. From this analysis, it was concluded that the biological function of leptins has changed significantly in the primates relative to the function of the leptin in the common ancestor of primates and rodents. This conclusion has several implications of importance, not the least being for pharmaceutical companies asked whether they should explore leptins as a pharmaceutical target. At the very least, it suggests that the mouse is not a good pharmacological model for compounds to be tested for their ability to combat obesity in humans. The post-genomic analysis suggests that a primate model must be used to test those compounds, with implications for the cost of developing an anti-obesity drug based on the leptin protein.

Intriguingly, a tree can also be built for the leptin receptor. Here, the evolutionary history is not so complete. In particular, fewer primate sequences are available for the leptin receptor than for leptin itself. Thus, the reconstructed ancestral sequences are less precise with the leptin receptor family, and the assignment of expressed and silent mutations to the tree are less certain. Nevertheless, it appears that the leptin receptor has undergone an episode of rapid sequence evolution in the primate half of the family as well. The example illustrates how much sequence data is needed (much) to build reliable models of this nature, as the ambiguity in the assignment of ancestral sequences makes it possible that the receptor was evolving rapidly not only in the lineage leading to primates but also in the lineage leading to mouse.

Nevertheless, the approximate correlation between the episode of rapid sequence evolution in the leptin family and in the leptin receptor family suggests a tool that might become useful in the advanced stages of post-genomic science when evolutionary histories are very well articulated. Here, it might be possible to detect ligand-receptor relationships between protein families in the database by a correspondence between their episodes of rapid sequence evolution. Thus, ligand families should evolve rapidly (in a non-Markovian fashion) at the same time in geological history as their receptors evolve. It will be interesting to identify more sequences for primate leptin receptors to see if a more complete evolutionary history allows us to see more clearly the co-evolution of the leptin receptor and leptin itself.

Example 5 Alcohol dehydrogenase

Mammalian alcohol dehydrogenase (E.C.1.1.1.1) have undergone a rapid episode of sequence evolution in and around the active site as substrate specificity has divergently evolved to handle xenobiotic substances in the liver. In contrast, over a comparable span of evolutionary distance, the active site of yeast alcohol dehydrogenase has changed very little, corresponding to an apparently constant role of the enzyme to act on the ethanol-acetaldehyde redox couple. Indeed, by identifying positions in mammalian dehydrogenases where amino acid variation was observed over a span of evolution where the same residues were conserved in the yeast dehydrogenases provided a clear map of the active site of the protein.

Example 6 Notch protein

A set of Notch homologs were obtained, and used to buid a multiple sequence alignment, and evolutionary tree (Drawing 6) and reconsructed intermediates throughout the evolutionary tree.

The functional interpretation based on these tools proceeded as follows. First, the f_2 values showed that the silent substitutions were not equilibrated over much of the tree. However, the f_2 value becomes close to 0.5 at points where the phyla diverge, suggesting near equilibration in the silent values. This defines the root of the tree near node 13. K_a/K_s values are given on the branches (numbers in italics). They suggest at the level of hypothesis that notch 1, notch 3 and notch are proteins with derived functions, while notch 4 is the paralog in mammals with the ancestral function. The rate constant for silent substitution is calculated to be ca. 23 x 10^{-9} changes/base per hear. This suggests that the notch paralogs diverged ca. 400 MYA. This is at the time of the development of advanced organis in vertebrates, suggesting that the Notch paralogs with derived function in the vertebrates are important for this level of organogenesis.

Example 7. C. elegans paralogs

NED distances are especially useful when comparing paralogs. Here, we need not worry so much about codon bias (it has at least been uniform among paralogs at any instant in evolutionary history). For example, we used the Master Catalog to identify all families of paralogs in the genome of *C. elegans*. Ca. 1250 families of paralogs with four or more members is found. We separated the families into in various classes using NED dates.

(a) Families where duplications all occurred > 400 MYA
(b) Families where duplications all occurred < 100 MYA
(c) Families where duplications have been ongoing throughout the past 400 MY.
(d) Families with duplications in specific episodes.
(e) Families showing a history of duplication > 400 MYA, but also having more recent episodes of recruitment.

Table 2 presents data from just five of these 1250 families.

Number of nodes generating paralogs in indicated time					
MYA	**0-100**	**100-200**	**200-300**	**300-400**	**>400**
gprod_19987 Mariner transposase	**39**	**1**	**4**	**0**	**5**
gprod_31705 similar to reverse transcriptase	**6**	**0**	**0**	**0**	**0**
gprod_32709 Histone H2A	**11**	**3**	**0**	**0**	**1**
gprod_7894 No definition line	**5**	**2**	**0**	**0**	**2**
gprod_19811 Serine-threonine kinase.	**5**	**2**	**3**	**5**	**39**

If the reviewer is a biomedical scientist, the Table immediately suggests ideas. Consider the family annotated as a serine-threonine kinase. It has 145 members in the Master Catalog; 55 or these are from *elegans*. The kinases generated by the recent duplications cannot part of the basic developmental plan of *elegans*; this was established 500 MYA. This raises questions: What is it about the serine-threonine kinases that recently diverged that might have something to do with recently evolved physiology? We then examine the K_a/K_s value within the Master Catalog trees, all with a click of a mouse button. We hypothesize which descendants of recent duplications performing the derived function, and which perform the primitive function. Dating the divergence, we try to make statements about changes in nematode biology that might be associated with the duplication. These hypotheses can now be tested by experiment (knock-outs, in particular).

	A	B	C	D	E	F		
family	0-	0.5-	1.0-	1.5-	2.0-	sum	average	description
	0.5	1.0	1.5	2.0	2.5		#char	
gprod_1025	0	0	0	0	5	5	143.4	
gprod_1063	0	0	1	0	2	3	46	
gprod_1069	0	0	1	0	2	3	3	
gprod_10729	1	0	0	0	3	4	143.5	
gprod_10751	0	0	1	0	5	6	204.667	
gprod_1090	0	0	0	0	3	3	35.3333	

gprod_1110	0	0	1	2	1	4	151.75	
gprod_1129	0	0	0	0	3	3	89.6667	
gprod_11679	0	0	0	0	3	3	48.3333	
gprod_1240	0	0	0	0	4	4	146	
gprod_12669	1	0	0	0	2	3	35	
gprod_1273	1	0	0	0	3	4	153	
gprod_12815	0	0	0	0	3	3	36	
gprod_1318	0	0	1	1	1	3	59.6667	
gprod_13259	0	0	1	1	6	8	55.875	
gprod_1354	0	0	0	0	3	3	152	
gprod_13591	0	0	0	0	5	5	116.8	
gprod_1405	0	0	0	1	1	2	222	
gprod_14189	0	0	2	0	14	16	601.625	similar to tubulin alpha-2 chain
gprod_1468	0	0	0	0	2	2	18	
gprod_1471	0	0	0	0	2	2	95	
gprod_15094	0	0	0	0	4	4	56	adenylate kinase
gprod_15198	0	0	1	0	2	3	41.6667	4-nitrophenylphosphatases
gprod_15375	0	0	0	0	5	5	42.8	Ce mostly, similar to triacylgycerol lipases
gprod_15390	1	0	0	3	13	17	916.529	guanylate cyclases, cat. domain protein kinases
gprod_15452	0	0	0	1	11	12	616.75	weak similarity to nodulation protein X
gprod_15464	2	0	0	0	5	7	242.429	serine protease inhibitor
gprod_15559	0	0	0	0	2	2	272	clathrin coat assembly like protein
gprod_15565	0	0	1	0	5	6	68	
gprod_15577	0	0	0	0	7	7	37	
gprod_15586	1	0	0	0	2	3	35.3333	
gprod_15588	0	0	1	0	4	5	135.2	
gprod_15724	1	0	0	0	2	3	60.3333	
gprod_15801	0	0	0	1	2	3	157.667	elongation factor EF
gprod_15805	0	0	0	2	3	5	330.4	
gprod_15819	0	0	2	0	5	7	169	putative integral membrane transport protein
gprod_15877	0	0	1	1	1	3	11.6667	
gprod_15878	2	0	0	1	0	3	1032	glyceraldehyde 3-phosphate dehydrogenase
gprod_15899	0	0	0	0	3	3	61	
gprod_15929	0	0	3	2	5	10	108.5	
gprod_15937	0	0	0	1	2	3	31	
gprod_15971	0	0	1	0	0	1	216	
gprod_15974	1	0	1	0	2	4	29.5	
gprod_16058	2	0	0	0	0	2	651.5	
gprod_16059	0	0	0	1	1	2	121	
gprod_16306	4	0	0	0	0	4	1220	
gprod_16402	1	0	0	0	4	5	81.4	
gprod_16477	0	0	0	0	0	0	0	
gprod_16653	0	0	0	1	2	3	73.6667	
gprod_16715	0	0	0	0	3	3	81	
gprod_1689	0	0	1	0	3	4	102.25	
gprod_16897	0	0	0	0	4	4	80.25	
gprod_16898	0	0	1	1	3	5	40.2	
gprod_17379	2	0	1	0	0	3	502	
gprod_1740	0	0	0	1	3	4	91.75	
gprod_17677	0	0	0	0	4	4	72.5	
gprod_1825	0	0	0	0	4	4	162.5	
gprod_19415	0	0	0	0	4	4	342	
gprod_19478	0	0	0	1	3	4	34.25	
gprod_19775	0	0	0	0	1	1	460	
gprod_19789	0	0	0	1	3	4	205.5	
gprod_19814	2	0	1	0	3	6	183.833	
gprod_19828	0	0	1	0	8	9	115.222	
gprod_19849	0	0	0	0	3	3	80.3333	
gprod_19867	2	0	1	1	12	16	227.625	

gprod_19899	0	0	1	2	4	7	100.143
gprod_19925	0	0	0	0	3	3	20.3333
gprod_19926	0	0	1	1	7	9	121.556
gprod_19931	0	0	0	0	0	0	0
gprod_19938	1	0	2	0	0	3	1151.33
gprod_19967	0	0	0	0	1	1	276
gprod_19971	4	0	0	0	3	7	418.571
gprod_19979	0	0	0	0	1	1	200
gprod_19983	0	0	0	0	7	7	116.429
gprod_20015	4	0	0	1	7	12	989.333
gprod_20024	0	0	1	0	2	3	61.3333
gprod_20031	1	0	2	0	2	5	39.6
gprod_20077	2	0	0	0	1	3	1050.67
gprod_20083	3	0	0	0	0	3	438
gprod_20110	0	0	0	0	1	1	212
gprod_20113	0	0	1	2	4	7	227.286
gprod_20115	0	0	0	0	6	6	213.833
gprod_20122	0	0	0	0	6	6	108.5
gprod_20124	1	0	3	0	9	13	95.3846
gprod_20125	0	0	1	0	2	3	79.3333
gprod_20126	0	0	4	0	7	11	62.8182
gprod_20154	0	0	0	1	5	6	138
gprod_20156	0	0	1	2	3	6	91
gprod_20169	0	0	4	4	11	19	939.842
gprod_20173	1	0	2	0	0	3	491.667
gprod_20188	0	0	1	1	7	9	95.8889
gprod_20196	0	0	1	0	7	8	197.5
gprod_20238	0	0	0	0	5	5	118.8
gprod_20244	0	0	0	0	3	3	82
gprod_20245	0	0	1	0	2	3	25.3333
gprod_20270	0	0	0	0	2	2	77
gprod_20295	1	0	1	0	7	9	274.111
gprod_20307	1	0	1	0	1	3	22
gprod_20367	0	0	0	0	2	2	97
gprod_20414	0	0	0	2	1	3	125.333
gprod_20418	0	0	1	0	2	3	154.333
gprod_20535	0	0	0	0	2	2	399
gprod_20557	0	0	0	0	1	1	140
gprod_20558	0	0	0	0	4	4	68.25
gprod_20569	1	0	0	1	1	3	110.333
gprod_20570	2	0	1	0	0	3	796.667
gprod_20576	0	0	2	0	4	6	294.667
gprod_20591	3	0	2	4	21	30	606.967
gprod_20624	0	0	0	0	1	1	210
gprod_20628	0	0	1	0	4	5	99.2
gprod_20641	0	0	2	3	6	11	533.091
gprod_20662	0	0	0	0	1	1	72
gprod_20664	0	0	1	2	2	5	30.4
gprod_20680	0	0	0	0	1	1	156
gprod_20700	2	0	0	0	1	3	73.3333
gprod_20727	4	0	0	1	3	8	399.125
gprod_20734	0	0	1	1	5	7	24.2857
gprod_20741	0	0	0	1	1	2	86
gprod_20765	0	0	1	0	7	8	70.375
gprod_20828	1	0	2	0	1	4	11
gprod_20829	0	0	0	1	1	2	132.5
gprod_20837	1	0	0	1	2	4	172.75
gprod_20852	0	0	0	1	2	3	151.333
gprod_20893	0	0	0	0	3	3	70.3333

gprod_20927	1	0	0	0	0	1	108
gprod_20938	0	0	3	0	0	3	85
gprod_21018	1	0	0	0	2	3	73.6667
gprod_21201	0	0	0	0	0	0	0
gprod_21246	0	0	0	0	3	3	62
gprod_21297	2	0	1	0	6	9	208.222
gprod_21313	0	0	0	0	1	1	260
gprod_21349	1	0	0	0	4	5	164.6
gprod_21518	0	0	0	0	5	5	128.4
gprod_21540	1	0	2	0	1	4	107.5
gprod_21543	2	0	0	0	3	5	232.6
gprod_21544	0	0	0	1	5	6	76.5
gprod_21553	1	0	0	0	0	1	678
gprod_21571	0	0	0	0	0	0	0
gprod_21635	1	0	0	0	0	1	1096
gprod_21693	1	0	0	0	1	2	88.5
gprod_2346	0	0	1	0	6	7	195.143
gprod_23932	3	0	1	0	4	8	170.25
gprod_24489	0	0	1	1	7	9	162.667
gprod_257	0	0	2	0	2	4	27.75
gprod_26542	0	0	0	0	0	0	0
gprod_26688	0	0	0	0	8	8	60.375
gprod_26772	0	0	1	2	10	13	262.615
gprod_26800	0	0	0	0	2	2	103
gprod_26933	0	0	0	0	1	1	62
gprod_26941	1	0	2	1	8	12	76.4167
gprod_27008	0	0	1	0	3	4	83.5
gprod_2725	0	0	2	0	0	2	338
gprod_27253	0	0	0	1	8	9	227
gprod_27327	1	0	1	0	1	3	81.3333
gprod_27505	0	0	0	0	1	1	60
gprod_27610	1	0	1	0	2	4	109.5
gprod_277	0	0	1	0	2	3	74.6667
gprod_27746	0	0	0	1	10	11	72.5455
gprod_279	0	0	0	0	1	1	129
gprod_28008	1	0	2	0	0	3	230.333
gprod_28099	0	0	1	0	4	5	66.6
gprod_28109	0	0	6	9	26	41	920.585
gprod_28114	0	0	0	0	2	2	102
gprod_28126	0	0	0	0	3	3	65.6667
gprod_28128	0	0	0	1	5	6	88.8333
gprod_28207	0	0	1	0	2	3	211.333
gprod_2836	0	0	0	1	10	11	161.909
gprod_29240	1	0	0	0	3	4	196.5
gprod_3136	0	0	0	0	4	4	70.5
gprod_31705	6	0	0	0	0	6	691.667
gprod_32138	0	0	0	0	6	6	42.5
gprod_32155	0	0	0	0	3	3	15.6667
gprod_32223	0	0	0	0	4	4	34.75
gprod_32385	0	0	0	0	2	2	29.5
gprod_32424	0	0	1	0	2	3	31
gprod_32450	0	0	0	0	5	5	24.8
gprod_3252	0	0	0	0	7	7	49.4286
gprod_32524	2	0	6	1	13	22	103.318
gprod_32586	0	0	0	1	0	1	304
gprod_32611	0	0	0	0	1	1	351
gprod_32623	0	0	0	0	5	5	49.8
gprod_32687	0	0	0	1	2	3	191.333
gprod_32711	0	0	0	2	7	9	260.889

gprod_32728	0	0	0	0	0	0	0
gprod_32739	0	0	0	0	4	4	465
gprod_32763	1	0	0	0	5	6	187
gprod_32765	0	0	1	0	3	4	53.25
gprod_32787	0	0	0	0	1	1	534
gprod_32798	0	0	0	1	5	6	172.667
gprod_32803	0	0	0	1	1	2	61
gprod_32817	1	0	0	0	2	3	129.667
gprod_32831	0	0	0	0	3	3	59.6667
gprod_32853	0	0	0	0	0	0	0
gprod_32854	0	0	1	2	6	9	115.778
gprod_3917	0	0	1	0	9	10	28.9
gprod_474	0	0	0	0	4	4	129.75
gprod_5001	0	0	0	1	3	4	39
gprod_5270	0	0	0	1	9	10	176.7
gprod_5276	0	0	1	0	2	3	17.6667
gprod_53	0	0	0	0	4	4	67.25
gprod_5760	0	0	0	2	3	5	71.6
gprod_5851	0	0	1	0	5	6	50
gprod_5932	0	0	1	0	1	2	61
gprod_5942	0	0	1	1	13	15	110
gprod_6132	0	0	0	2	5	7	135.857
gprod_6992	0	0	0	0	3	3	41.6667
gprod_7278	0	0	0	0	8	8	190.75
gprod_746	2	0	1	1	0	4	116
gprod_747	0	0	0	0	6	6	129.167
gprod_7647	0	0	0	0	2	2	175.5
gprod_7650	0	0	0	1	3	4	425.75
gprod_7655	0	0	2	3	1	6	68
gprod_7658	0	0	0	1	2	3	29.6667
gprod_7670	3	0	0	3	10	16	241.188
gprod_7675	3	0	0	0	0	3	175.667
gprod_7683	0	0	0	1	2	3	57.3333
gprod_7688	0	0	1	1	2	4	87
gprod_7696	0	0	0	1	3	4	35.25
gprod_7701	2	0	0	0	5	7	465.571
gprod_7706	1	0	0	3	1	5	117.6
gprod_7714	0	0	0	0	5	5	188.8
gprod_7715	0	0	0	0	2	2	91
gprod_7731	2	0	3	0	0	5	1163.8
gprod_7733	0	0	1	4	29	34	339.382
gprod_7735	0	0	0	0	3	3	54.6667
gprod_7739	0	0	1	2	3	6	57.3333
gprod_7743	1	0	0	0	3	4	97.75
gprod_7744	0	0	3	0	10	13	166.462
gprod_7764	0	0	0	0	11	11	157.182
gprod_7766	0	0	0	1	5	6	101.167
gprod_7770	0	0	0	0	4	4	74.5
gprod_7773	0	0	1	0	3	4	38.75
gprod_7778	0	0	0	0	1	1	88
gprod_7779	0	0	0	0	1	1	184
gprod_7783	0	0	1	0	3	4	87.25
gprod_7800	2	0	0	0	3	5	129.2
gprod_7809	0	0	0	0	3	3	95.3333
gprod_7816	0	0	1	0	1	2	60.5
gprod_7818	1	0	0	0	3	4	71
gprod_7838	0	0	0	0	5	5	40.4
gprod_7852	0	0	0	0	4	4	78
gprod_7853	0	0	0	1	6	7	58.4286

gprod_7856	3	0	0	0	0	3	334
gprod_7863	0	0	0	0	3	3	48.6667
gprod_7866	0	0	1	0	3	4	143.5
gprod_7880	0	0	0	0	3	3	29.3333
gprod_7882	0	0	2	0	2	4	48.5
gprod_7890	0	0	0	1	11	12	87.0833
gprod_7891	0	0	0	3	13	16	137.25
gprod_7909	0	0	0	0	5	5	35.4
gprod_7932	1	0	1	0	1	3	129.667
gprod_7938	0	0	0	1	2	3	18
gprod_7955	1	0	0	1	5	7	379.286
gprod_7956	0	0	0	0	9	9	576.667
gprod_7964	0	0	1	1	2	4	85
gprod_7970	3	0	2	1	3	9	127.111
gprod_7978	0	0	1	0	3	4	40.5
gprod_7980	0	0	0	0	6	6	59.6667
gprod_7989	1	0	0	0	4	5	703.2
gprod_7997	0	0	0	1	7	8	81.5
gprod_8011	0	0	1	0	2	3	41
gprod_8019	0	0	0	1	3	4	102.25
gprod_8021	1	0	2	0	4	7	200.286
gprod_8023	0	0	2	2	9	13	236.615
gprod_8032	0	0	0	0	2	2	78
gprod_8035	2	0	0	1	0	3	308.667
gprod_8046	0	0	0	1	2	3	31
gprod_8048	0	0	2	2	22	26	648.5
gprod_8069	0	0	1	1	1	3	85.6667
gprod_8072	0	0	0	0	3	3	123.667
gprod_8073	0	0	1	1	9	11	174.636
gprod_8080	0	0	1	1	3	5	152
gprod_8083	0	0	0	0	3	3	96.6667
gprod_8094	1	0	0	0	3	4	72.75
gprod_8095	0	0	0	0	3	3	125.667
gprod_8096	0	0	0	0	3	3	126.667
gprod_8097	0	0	0	0	6	6	173.167
gprod_8106	0	0	1	0	4	5	186
gprod_8114	1	0	0	0	2	3	35.6667
gprod_8115	0	0	0	0	10	10	161.1
gprod_8116	1	0	1	0	2	4	182
gprod_8119	0	0	1	0	3	4	101.5
gprod_8132	0	0	0	0	4	4	152.25
gprod_8138	0	0	1	1	1	3	96.6667
gprod_814	0	0	0	0	3	3	74
gprod_8140	1	0	1	1	9	12	299.083
gprod_815	0	0	0	1	2	3	88
gprod_8170	1	0	0	0	2	3	251
gprod_8188	1	0	0	0	4	5	81.2
gprod_8199	1	0	1	0	8	10	294.7
gprod_8245	0	0	1	0	5	6	85.3333
gprod_8270	0	0	1	0	0	1	504
gprod_8288	0	0	0	0	2	2	41
gprod_8289	0	0	1	0	2	3	54
gprod_8300	0	0	2	0	2	4	116.25
gprod_8313	1	0	0	0	2	3	107.333
gprod_8335	1	0	2	0	0	3	236.667
gprod_8341	1	0	1	3	32	37	584.351
gprod_8355	0	0	0	0	2	2	128.5
gprod_8361	0	0	0	0	3	3	9
gprod_8384	0	0	1	0	2	3	64.6667

gprod_8424	0	0	0	0	5	5	102.2
gprod_8433	1	0	0	0	1	2	52.5
gprod_8439	0	0	0	1	7	8	270.875
gprod_8463	0	0	0	0	0	0	0
gprod_8465	1	0	0	0	3	4	45.25
gprod_8470	0	0	0	1	7	8	165.25
gprod_8485	0	0	0	0	2	2	56
gprod_8495	1	0	2	1	5	9	60.6667
gprod_8500	1	0	0	1	0	2	79.5
gprod_8511	2	0	1	0	0	3	178.333
gprod_8538	0	0	1	0	2	3	72
gprod_8568	0	0	1	0	2	3	65
gprod_8574	1	0	2	0	3	6	158.833
gprod_8576	0	0	0	0	3	3	69.6667
gprod_8585	0	0	1	0	11	12	187.333
gprod_8603	1	0	0	0	4	5	138.6
gprod_8610	0	0	0	2	10	12	80.9167
gprod_8614	0	0	0	0	4	4	121.5
gprod_8619	0	0	1	1	2	4	68.25
gprod_8620	0	0	0	0	3	3	122.667
gprod_8643	0	0	0	1	2	3	148.667
gprod_8650	0	0	1	0	2	3	402
gprod_8659	0	0	0	1	4	5	304.6
gprod_8669	0	0	1	0	1	2	67
gprod_8684	1	0	2	0	0	3	97
gprod_8690	0	0	0	0	3	3	76
gprod_8695	0	0	0	0	1	1	186
gprod_8711	2	0	0	0	2	4	281
gprod_8727	0	0	1	0	2	3	217.333
gprod_8737	1	0	1	0	1	3	104.333
gprod_8750	1	0	0	0	11	12	138.333
gprod_8758	0	0	0	0	3	3	26.6667
gprod_8768	0	0	1	1	4	6	537.167
gprod_8809	3	0	0	0	0	3	32.3333
gprod_8810	0	0	0	1	2	3	70.6667
gprod_8825	0	0	1	1	2	4	78.5
gprod_8829	1	0	2	0	0	3	707
gprod_8838	0	0	0	0	3	3	28.6667
gprod_885	0	0	0	0	3	3	54
gprod_8863	0	0	0	0	3	3	68.3333
gprod_8881	0	0	0	0	3	3	87.3333
gprod_8914	0	0	0	0	3	3	35.3333
gprod_8950	0	0	0	0	3	3	44
gprod_8984	0	0	0	0	0	0	0
gprod_9042	0	0	0	1	2	3	62
gprod_9052	0	0	0	0	3	3	86
gprod_9053	0	0	0	2	1	3	68
gprod_9086	0	0	1	0	9	10	78.6
gprod_9087	0	0	1	1	5	7	76.2857
gprod_909	0	0	2	1	14	17	177
gprod_9112	0	0	0	1	3	4	89.75
gprod_9125	0	0	0	0	1	1	96
gprod_9146	0	0	1	0	6	7	25.1429
gprod_9196	0	0	0	0	3	3	51.6667
gprod_9247	0	0	0	1	1	2	148.5
gprod_9291	0	0	0	0	1	1	60
gprod_9300	3	0	0	0	0	3	280.333
gprod_9337	0	0	0	0	3	3	47.6667
gprod_934	0	0	0	0	2	2	39.5

gprod_9365	2	0	0	0	0	2	62.5
gprod_9379	0	0	0	0	1	1	309
gprod_9506	0	0	1	2	1	4	168.75
gprod_9545	2	0	0	0	1	3	59.3333
gprod_9582	0	0	0	0	0	0	0
gprod_963	0	0	0	0	3	3	69
gprod_9814	0	0	0	0	2	2	121
gprod_12760	0	1	0	1	1	3	136.667
gprod_12786	6	1	1	0	11	19	264.053
gprod_13977	0	1	1	1	6	9	54.8889
gprod_14170	2	1	0	5	11	19	297.579
gprod_1425	0	1	0	1	4	6	123
gprod_14535	1	1	3	2	15	22	170.318
gprod_14642	0	1	1	0	12	14	1276.43
gprod_14919	0	1	1	2	5	9	337.444
gprod_16029	3	1	0	1	0	5	62.6
gprod_16115	0	1	0	1	3	5	80.4
gprod_16281	1	1	1	0	0	3	58.6667
gprod_1644	2	1	0	0	1	4	195.25
gprod_16975	0	1	0	0	3	4	70.75
gprod_1927	0	1	0	0	5	6	290.333
gprod_1934	0	1	3	0	8	12	184.417
gprod_19774	0	1	1	0	5	7	89
gprod_19796	1	1	1	2	14	19	467.368
gprod_19831	0	1	1	2	11	15	215.133
gprod_19846	3	1	2	0	25	31	700.161
gprod_19848	0	1	0	0	3	4	87
gprod_19877	0	1	0	0	2	3	48.6667
gprod_19893	3	1	1	1	14	20	812.95
gprod_19895	2	1	4	0	4	11	462.545
gprod_19900	0	1	2	1	7	11	61.5455
gprod_19903	0	1	1	0	7	9	246.556
gprod_19914	0	1	0	0	2	3	33.6667
gprod_19924	1	1	1	0	14	17	492.118
gprod_19930	0	1	0	1	3	5	114.8
gprod_19941	0	1	0	1	2	4	37.25
gprod_19981	1	1	6	6	35	49	1243.14
gprod_19986	1	1	4	5	14	25	543.2
gprod_20003	0	1	0	1	5	7	115
gprod_20020	0	1	0	0	0	1	126
gprod_20030	0	1	0	0	8	9	69.2222
gprod_20049	0	1	0	0	7	8	129.125
gprod_20149	0	1	1	0	1	3	53.3333
gprod_20159	0	1	0	1	7	9	252
gprod_20261	2	1	0	0	0	3	123
gprod_20304	0	1	0	0	3	4	219.75
gprod_20533	2	1	0	0	1	4	156.5
gprod_20554	0	1	1	1	0	3	251.667
gprod_20585	0	1	1	1	1	4	60
gprod_20633	0	1	0	2	3	6	177
gprod_2064	0	1	0	1	2	4	138
gprod_20642	0	1	0	0	3	4	188
gprod_20682	0	1	0	0	1	2	42
gprod_20753	0	1	0	0	2	3	1571.67
gprod_20784	0	1	0	2	1	4	192
gprod_20935	0	1	0	0	3	4	57
gprod_20995	0	1	1	1	1	4	33.5
gprod_21290	0	1	3	2	2	8	176.625
gprod_21305	0	1	0	0	1	2	58.5

gprod_21306	0	1	1	0	1	3	211.667
gprod_21527	0	1	2	0	7	10	374.4
gprod_21528	0	1	2	1	1	5	49.2
gprod_23639	4	1	1	0	7	13	341.462
gprod_23919	2	1	1	1	4	9	73.5556
gprod_2606	1	1	2	0	0	4	67.5
gprod_26860	0	1	0	0	10	11	115
gprod_26964	0	1	0	0	1	2	7.5
gprod_27333	2	1	0	0	3	6	313.333
gprod_27592	0	1	0	0	3	4	65
gprod_28106	0	1	1	0	4	6	37.3333
gprod_28112	1	1	0	0	7	9	210.778
gprod_28113	0	1	0	2	1	4	133.25
gprod_28129	0	1	0	0	1	2	38
gprod_2863	0	1	0	0	11	12	96.9167
gprod_31702	2	1	0	0	0	3	34
gprod_31741	2	1	0	0	1	4	23
gprod_31781	0	1	0	0	1	2	34
gprod_32030	2	1	2	2	10	17	419.588
gprod_32193	0	1	2	1	1	5	216
gprod_32422	0	1	1	0	3	5	42
gprod_32443	0	1	0	4	4	9	269.333
gprod_32714	1	1	3	1	3	9	81.5556
gprod_32721	2	1	0	0	14	17	510.059
gprod_32722	0	1	1	3	18	23	188
gprod_32723	1	1	0	0	1	3	118.333
gprod_32725	0	1	0	2	10	13	349.308
gprod_32741	1	1	2	0	3	7	31.7143
gprod_32752	0	1	0	0	3	4	83.5
gprod_32786	0	1	0	1	7	9	217.778
gprod_32809	0	1	1	2	4	8	44.75
gprod_32834	1	1	0	0	0	2	119
gprod_32855	0	1	1	1	6	9	103.333
gprod_4443	0	1	1	3	3	8	40.875
gprod_470	1	1	0	0	0	2	605.5
gprod_5522	1	1	1	1	1	5	54
gprod_564	0	1	0	0	3	4	68.25
gprod_5670	0	1	0	1	1	3	47.6667
gprod_6233	0	1	0	0	10	11	287.091
gprod_720	0	1	0	0	0	1	42
gprod_7653	0	1	1	1	1	4	168.5
gprod_7656	0	1	0	0	5	6	70.3333
gprod_7678	1	1	1	2	0	5	85.6
gprod_7681	0	1	1	0	5	7	229.286
gprod_7686	0	1	0	0	2	3	56
gprod_7697	0	1	1	0	7	9	249
gprod_7702	0	1	0	1	8	10	138.4
gprod_7749	0	1	0	1	4	6	263.333
gprod_7777	0	1	0	1	4	6	74.3333
gprod_7793	0	1	0	2	3	6	75.1667
gprod_7797	0	1	0	1	4	6	74.1667
gprod_7821	1	1	0	0	1	3	224
gprod_7825	0	1	0	0	3	4	114.25
gprod_7829	0	1	1	3	15	20	483.6
gprod_7841	0	1	1	0	1	3	148.333
gprod_7844	0	1	0	1	14	16	287.875
gprod_7848	1	1	0	1	5	8	124.375
gprod_7870	0	1	0	3	1	5	153.6
gprod_7878	0	1	0	0	2	3	14.3333

gprod_7906	0	1	2	0	10	13	301.538
gprod_7908	0	1	2	1	6	10	343.4
gprod_7933	0	1	3	0	9	13	310.462
gprod_7984	1	1	2	1	1	6	53.1667
gprod_7985	0	1	0	2	3	6	108
gprod_8005	0	1	0	0	2	3	25.6667
gprod_8010	1	1	0	0	1	3	82
gprod_8028	1	1	1	1	14	18	353.333
gprod_8038	1	1	0	2	1	5	60.8
gprod_8054	2	1	0	1	5	9	523.556
gprod_8055	3	1	2	0	9	15	294.267
gprod_8065	0	1	0	2	5	8	154.5
gprod_8074	0	1	0	1	1	3	81.3333
gprod_8092	0	1	1	1	2	5	132.8
gprod_8098	1	1	0	0	1	3	117.667
gprod_8113	0	1	2	0	1	4	50.75
gprod_8122	0	1	0	1	2	4	199.5
gprod_8301	1	1	1	1	3	7	136.714
gprod_8311	0	1	1	0	2	4	98.75
gprod_8316	0	1	0	0	2	3	28.6667
gprod_8321	3	1	0	1	2	7	103.571
gprod_8334	0	1	0	0	3	4	72.25
gprod_8359	3	1	1	0	6	11	87.2727
gprod_8438	0	1	1	0	4	6	215
gprod_8449	2	1	0	0	0	3	81.6667
gprod_8475	1	1	1	0	2	5	301.8
gprod_8499	0	1	0	0	2	3	107.667
gprod_8554	1	1	1	0	6	9	88.4444
gprod_8595	0	1	2	1	0	4	417.5
gprod_8599	0	1	1	2	7	11	142.273
gprod_8600	1	1	0	1	11	14	149.5
gprod_8608	0	1	1	0	3	5	41.6
gprod_8671	0	1	0	1	4	6	88.6667
gprod_8732	1	1	0	0	3	5	343.8
gprod_8733	0	1	0	0	2	3	33.3333
gprod_8742	0	1	0	0	4	5	104.8
gprod_8743	0	1	1	0	1	3	36.3333
gprod_8767	0	1	1	0	3	5	163.8
gprod_8780	0	1	4	2	12	19	359.263
gprod_8860	1	1	1	3	1	7	103.714
gprod_9011	0	1	1	0	14	16	826.312
gprod_9024	0	1	1	0	3	5	6.2
gprod_9160	0	1	0	1	4	6	55.6667
gprod_9271	1	1	2	0	2	6	81.3333
gprod_9377	0	1	0	0	5	6	66
gprod_9404	0	1	0	0	2	3	31.6667
gprod_1191	0	2	0	1	0	3	91.6667
gprod_182	1	2	1	0	15	19	336.789
gprod_18780	1	2	1	1	9	14	815
gprod_19811	4	2	1	2	45	54	833.111
gprod_19819	2	2	1	0	0	5	221.8
gprod_19837	0	2	0	1	4	7	430.429
gprod_19840	1	2	0	0	0	3	269.667
gprod_19891	1	2	1	1	7	12	251.5
gprod_19946	0	2	2	5	9	18	299.5
gprod_19952	1	2	4	2	18	27	520.037
gprod_19953	0	2	2	4	12	20	343.85
gprod_19969	0	2	0	1	6	9	251.556
gprod_19987	38	2	1	3	5	49	2589.8

gprod_20036	2	2	0	0	1	5	60.8	
gprod_20062	1	2	1	0	7	11	168	
gprod_20136	2	2	3	2	4	13	166.692	
gprod_20175	0	2	4	2	7	15	320.867	
gprod_20214	0	2	2	3	20	27	371.296	
gprod_20555	1	2	0	0	0	3	930.333	
gprod_20562	0	2	0	1	1	4	186.5	
gprod_20716	1	2	0	0	1	4	2526.5	
gprod_20718	0	2	1	1	2	6	54.6667	
gprod_21454	2	2	2	0	8	14	217.643	
gprod_21524	5	2	2	0	3	12	1030	
gprod_23703	0	2	0	0	1	3	43.6667	
gprod_26509	2	2	0	2	0	6	158.167	
gprod_26822	1	2	0	0	5	8	157.375	
gprod_28009	0	2	0	0	6	8	169.875	
gprod_30860	0	2	1	1	13	17	540.706	
gprod_32324	1	2	0	0	5	8	84.875	
gprod_32353	1	2	2	1	4	10	92.3	
gprod_32389	1	2	0	1	4	8	105.75	
gprod_32514	3	2	5	3	17	30	78.1333	
gprod_32663	2	2	1	2	24	31	653.065	
gprod_32766	2	2	0	0	2	6	83.5	
gprod_32802	3	2	0	0	1	6	156.667	
gprod_32806	1	2	1	0	0	4	70.5	
gprod_32878	0	2	0	0	8	10	76	
gprod_7742	0	2	1	0	2	5	220	
gprod_7745	0	2	3	0	11	16	167.688	
gprod_7760	0	2	3	0	2	7	60.5714	
gprod_7869	0	2	0	1	10	13	90	
gprod_7884	1	2	4	0	5	12	225	
gprod_8008	1	2	4	2	15	24	213.25	
gprod_8064	0	2	1	3	11	17	356.176	
gprod_8100	0	2	0	0	0	2	46	
gprod_8130	0	2	1	0	0	3	70.6667	
gprod_8131	1	2	0	0	2	5	298.4	
gprod_8352	1	2	9	19	22	53	1635.38	
gprod_8365	2	2	0	0	1	5	68.8	
gprod_8440	0	2	3	1	3	9	131.556	
gprod_8635	1	2	0	0	0	3	425.333	
gprod_8876	1	2	1	0	1	5	41.2	
gprod_9040	0	2	0	1	0	3	55	
gprod_916	1	2	0	1	4	8	50.75	
gprod_14386	4	3	6	2	29	44	577.295	
gprod_19868	0	3	1	0	3	7	347.714	
gprod_19999	1	3	3	0	4	11	241.364	
gprod_20011	0	3	1	0	6	10	52.3	
gprod_20260	0	3	0	1	0	4	96.5	
gprod_20542	1	3	3	2	14	23	694.565	
gprod_20669	3	3	4	2	9	21	307.286	
gprod_28110	3	3	2	3	5	16	452.938	
gprod_32423	2	3	0	1	11	17	194.941	
gprod_32527	0	3	2	2	4	11	74.6364	
gprod_32795	1	3	2	2	2	10	85.4	
gprod_5572	1	3	0	0	1	5	27.8	
gprod_7741	0	3	3	1	7	14	60.4286	
gprod_7828	2	3	5	4	25	39	673.872	
gprod_7894	3	3	1	0	2	9	169.889	
gprod_15348 polyproteins	8	4	9	3	18	42	429.881	Ce only weak sim to retrovirus-related

gprod_19947	4	4	6	4	19	37	444.676	sim to C. elegans olfactory receptor ODR-10
gprod_20160	1	4	5	3	17	30	357.4	similar to G-protein coupled receptor
gprod_20201	4	4	4	3	13	28	397.5	Ce only, LET-23 receptor protein-tyrosine kinase
gprod_24016	4	4	4	3	22	37	264.73	Ce olfactory receptor ODR-10
gprod_26736	4	4	6	4	7	25	327.36	chitinase, also ascomycetes
gprod_31697	3	4	9	8	47	71	734.211	**C with other Rhab** collagen
gprod_32713	5	4	6	5	17	37	597.108	**Ce only**
gprod_32730	6	4	4	3	13	30	205.633	**Ce only**
gprod_7689	0	4	4	6	21	35	496.486	Celegans only
gprod_7690	5	4	1	0	0	10	294	Histone H2B
gprod_7868	1	4	4	3	11	23	279.043	chitinase
gprod_8077	1	4	1	3	7	16	201.25	Ce only
gprod_20981	5	5	5	4	18	37	486.946	Ce only similar to collagen
gprod_32709	7	5	2	0	1	15	218.8	Histone
gprod_7769	0	5	0	0	1	6	111.167	Ce only
gprod_7836	2	5	5	5	25	42	295.452	similar to Lectin C-type domain
gprod_9070	2	5	6	1	4	18	171.611	**CE only**
gprod_15391	3	6	5	8	17	39	634.103	Ce only weak sim mouse Zn finger 5 protien
gprod_20153	2	6	10	4	26	48	728.292	Ce only
gprod_21539	3	6	0	0	0	9	573.111	Ce only
gprod_7758	4	6	2	1	7	20	284.45	Histone H3
gprod_8309	5	6	5	2	16	34	243.588	mariner transposase
gprod_15294	4	7	3	0	19	33	1696.73	Ce only
gprod_2547	1	7	0	1	0	9	189.111	histone
gprod_7771	0	7	3	4	14	28	204.321	Ce alone
gprod_8784	7	7	2	1	11	28	479.179	Ce alone
gprod_19851	13	10	15	15	22	75	1031.37	Ce only
gprod_19882	5	11	13	6	38	73	827.479	Ce only similar to Transposase
gprod_14141	5	12	9	14	143	183	2326.21	Ce only
gprod_19800	15	13	7	14	88	137	1154.35	Ce only
gprod_26731	21	14	17	14	73	139	864.892	Ce only olfactory receptor ODR-10
gprod_32715	5	16	4	2	4	31	807.806	major sperm protein

One observation apparent from the Table is that genes that have multiple recent recruitments in *C. elegans* are unlikely to have clearly identifiable homologs in other phyla, while those that have few recent recruitments are more likely than average to have clearly identifiable homologs in other phyla.